

Flores Saldivar, Alfredo Alan (2018) Predicting potential customer needs and wants for agile design and manufacture in an industry 4.0 environment. PhD thesis.

<https://theses.gla.ac.uk/38974/>

Copyright and moral rights for this work are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This work cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Predicting Potential Customer Needs and Wants for Agile Design and Manufacture in an Industry 4.0 Environment

by

Alfredo Alan Flores Saldivar

B. S Industrial and Systems Engineering, Autonomous University of the
North East, Saltillo, Coahuila, Mexico, 2011

M. S Advanced Manufacturing Systems, Mexican Corporation of Materials
Research, Saltillo, Coahuila, Mexico, 2013

Thesis submitted in complete fulfilment of the requirements for the degree of
Doctor of Philosophy (Ph.D.) to the School of Engineering at the University of
Glasgow

Glasgow, United Kingdom.

Abstract

Manufacturing is currently experiencing a paradigm shift in the way that products are designed, produced and serviced. Such changes are brought about mainly by the extensive use of the Internet and digital technologies. As a result of this shift, a new industrial revolution is emerging, termed “Industry 4.0” (i4), which promises to accommodate mass customisation at a mass production cost. For i4 to become a reality, however, multiple challenges need to be addressed, highlighting the need for design for agile manufacturing and, for this, a framework capable of integrating big data analytics arising from the service end, business informatics through the manufacturing process, and artificial intelligence (AI) for the entire manufacturing value chain.

This thesis attempts to address these issues, with a focus on the need for design for agile manufacturing. First, the state of the art in this field of research is reviewed on combining cutting-edge technologies in digital manufacturing with big data analysed to support agile manufacturing. Then, the work is focused on developing an AI-based framework to address one of the customisation issues in smart design and agile manufacturing, that is, prediction of potential customer needs and wants.

With this framework, an AI-based approach is developed to predict design attributes that would help manufacturers to decide the best virtual designs to meet emerging customer needs and wants predictively. In particular, various machine learning approaches are developed to help explain at least 85% of the design variance when building a model to predict potential customer needs and wants. These approaches include k-means clustering, self-organizing maps, fuzzy k-means clustering, and decision trees, all supporting a vector machine to evaluate and extract conscious and subconscious customer needs and wants. A model capable of accurately predicting customer needs and wants for at least 85% of classified design attributes is thus

obtained. Further, an analysis capable of determining the best design attributes and features for predicting customer needs and wants is also achieved.

As the information analysed can be utilized to advise the selection of desired attributes, it is fed back in a closed-loop of the manufacturing value chain: design → manufacture → management/service →→→ design... For this, a total of 4 case studies are undertaken to test and demonstrate the efficacy and effectiveness of the framework developed. These case studies include: 1) an evaluation model of consumer cars with multiple attributes including categorical and numerical ones; 2) specifications of automotive vehicles in terms of various characteristics including categorical and numerical instances; 3) fuel consumptions of various car models and makes, taking into account a desire for low fuel costs and low CO₂ emissions; and 4) computer parts design for recommending the best design attributes when buying a computer. The results show that the decision trees, as a machine learning approach, work best in predicting customer needs and wants for smart design.

With the tested framework and methodology, this thesis overall presents a holistic attempt to addressing the missing gap between manufacture and customisation, that is meeting customer needs and wants. Effective ways of achieving customization for i4 and smart manufacturing are identified. This is achieved through predicting potential customer needs and wants and applying the prediction at the product design stage for agile manufacturing to meet individual requirements at a mass production cost. Such agility is one key element in realising Industry 4.0. At the end, this thesis contributes to improving the process of analysing the data to predict potential customer needs and wants to be used as inputs to customizing product designs agilely.

Table of Contents

Abstract	i
Table of Contents.....	iii
List of Tables	vii
List of Figures	viii
Preface	xii
Acknowledgements	xiii
Author's Declaration	xv
Chapter 1 Introduction.....	1
1.1 Industry 4.0	2
1.1.1 Why Industry 4.0 Is Important	4
1.1.2 Components of Industry 4.0	5
1.1.3 Smart Factory.....	10
1.2 Aims of This Research	11
1.3 Outline of the Thesis.....	15
Chapter 2 Literature Review	17
2.1 Smart Manufacturing Developments.....	17
2.2 Industry 4.0 - Mass Customization and the Entire Value Chain	20
2.2.1 Entire Value Chain	23
2.2.2 Gaps Between Current Manufacturing Systems and Industry 4.0.....	26
2.3 Cyber-Physical Integration Realising Smart Manufacturing	29
2.3.1 Cyber-Physical Integration	32
2.3.2 Embedded Manufacturing Systems.....	38
2.3.3 CPS and Data Analytics for Smart Manufacturing.....	39
2.4 Big Data and Business Informatics for Industry 4.0.....	40
2.4.1 Role of Big Data Analytics in IoT	43
2.4.2 Big Data Analytics Tools for the Smart Manufacturing Value Chain.....	46
2.5 Challenges Achieving Mass Customization	52
2.6 Summary	55
Chapter 3 Methods for Attribute Prediction Using Smart Design Principles.....	63
3.1 Hypotheses to Set the Scene	64

3.2 Artificial Intelligence for the Smart Manufacturing Value Chain	66
3.2.1 Predictive Models to Address Customer Needs.....	71
3.2.2 Affective Design for Mass Customization	74
3.3 Machine Learning Based Approaches	77
3.3.1 Clustering Analysis	79
3.3.2 Classification Analysis	85
3.3.3 Feature Selection Analysis	89
3.4 Smart Design Under Industry 4.0 Principles	91
3.4.1 Automated Design for Industry 4.0	92
3.4.2 Motivation of Selected approaches	95
3.4.3 Case Studies of Predicting Potential Customer Needs and Wants for Future CAutoD	99
3.5 Summary	103
Chapter 4 Framework for Predicting Potential Customer Needs and Wants	104
4.1 Value Chain for Predicting Potential Customer Needs and Wants	104
4.2 Artificial Intelligence for a Closed-Loop Framework.....	108
4.3 Classification Learner Framework for Coding Customer Needs.....	111
4.4 Genetic Search Framework for Selecting Best Attributes	112
4.5 Summary	118
Chapter 5 Applications and Case Studies.....	120
5.1 Datasets for Applications.....	120
5.1.1 Car Evaluation Dataset	121
5.1.2 Automobile Dataset	122
5.1.3 Fuel Economy Dataset	124
5.1.4 CPU Dataset.....	127
5.2 Selected Case Studies to Illustrate the Applications	130
5.3 Data Analysis	132
5.3.1 Car Evaluation Dataset Results.....	132
5.3.2 Automobile Dataset Results	138
5.3.3 Fuel Economy Dataset Results.....	143
5.3.4 CPU Dataset Results	158
5.4 Evaluation of the Cases as a Result of Machine Learning Approaches.....	175

5.4.1 Car Case Evaluation	175
5.4.2 Automobile Case Evaluation	176
5.4.3 Fuel Economy Case Evaluation	177
5.4.4 CPU Case Evaluation	178
5.5 Summary	178
Chapter 6 Conclusions and Future Work.....	181
6.1 General Conclusion.....	181
6.1.1 Car Evaluation Dataset	182
6.1.2 Automobile Dataset	183
6.1.3 Fuel Economy Dataset	183
6.1.4 CPU Dataset.....	184
6.2 Reflections on the Hypotheses	184
6.3 Future Directions	187
References.....	190

List of Tables

Table 1–1 Design principles of each Industry 4.0 component.....	7
Table 2–1 Existing technologies for big data analysis and machine learning	47
Table 2–2 Summary of the state of the art.....	56
Table 3–1 Digital manufacturing enabler technologies	68
Table 5–1 Car evaluation dataset [116]	121
Table 5–2 Automobile data	123
Table 5–3 Description of fuel economy data attribute.	125
Table 5–4 Description of CPU data attribute.....	128
Table 5–5 Confusion matrix for the SOM.....	133
Table 5–6 Model accuracy by class.....	133
Table 5–7 Simple k-means clustering testing 7 attributes.....	134
Table 5–8 Model accuracy by class.....	135
TABLE 5–9 Confusion matrix for classified attributes.....	135
Table 5–10 Feature selection results using genetic search	171
Table 5–11 Model accuracy evaluation of AI approaches for the car evaluation dataset	175
Table 5–12 Model accuracy evaluation of AI approaches for the automobile dataset	176
Table 5–13 Model accuracy evaluation of AI approaches for the fuel economy dataset.	177
Table 5–14 Model accuracy evaluation of AI approaches for the CPU dataset.	178

List of Figures

Figure 1-1 Trend to mass customization according to [9].	5
Figure 1-2 Industrial Revolutions and evolution of manufacturing towards Industry 4.0 [3].	6
Figure 2-1 A comparison of a Value Chain with a Supply Chain [33].	25
Figure 2-2 Research gap between recent manufacturing systems and i4 [39].	29
Figure 2-3 Integrated Approach to develop CPS -. [40].	30
Figure 2-4 Model Integration: OpenMETA framework. - [45].	33
Figure 2-5 Method Integration Framework. - [45].	35
Figure 2-6 Tool Integration Framework. - [45].	37
Figure 2-7 Architecture for implementing CPS [50]	40
Figure 2-8 Expected growth in digital data from 2010 to 2020 [59].	51
Figure 2-9 Horizontal and vertical integration in mass customization [60].	53
Figure 3-1 Framework for real-time decision informatics [69].	67
Figure 3-2 Basic architecture of a recommendation system [72].	72
Figure 3-3 Prediction process of product configuration for new customers [81].	77
Figure 3-4 Framework for the Feature Genetic Search [108].	91
Figure 3-5 CAutoD realised through an evolutionary computing process [26]	93
Figure 4-1 Value chain closed loop for predictive customer needs and wants.	107
Figure 4-2 Industry 4.0 automated closed-loop for predicting customer needs and wants for customization [27]	110
Figure 4-3 Proposed AI-based methodology for predictive data analysis and attribute classification.	111
Figure 4-4 Industry 4.0 closed-loop for predicting customer needs and wants using data mining approaches	115
Figure 4-5 Genetic search framework for feature selection.	117
Figure 5-1 Distributions of car evaluation dataset for customization.	122
Figure 5-2 Relation of fuel economy dataset for the average USD spent, the average CO ₂ emissions, and the average annual fuel costs of conventional fuel.	127
Figure 5-3 Relation of recommended customer price attribute vs product collection or models attribute.	130
Figure 5-4 Results of tested data. SOM weight distances on the left, and SOM clusters found on the right.	133
Figure 5-5 Scatter plot for the incorrect classified instances of variables “buy price vs “doors” using simple k-means.	136
Figure 5-6 Scatter plot for the incorrect classified instances of variables “buy price vs “repair price (maintenance)” using simple k-means.	137
Figure 5-7 Results of tested data. Fuzzy c-means with 3 clusters found.	139
Figure 5-8 Membership function. From top to bottom: cluster 1, 2 and 3 results.	141
Figure 5-9 Confusion matrix obtained for positive predictive values.	142
Figure 5-10 Parallel coordinates plot for membership functions.	143

Figure 5-11 Scatter plot for the correct instances using the decision trees classifiers of variable “spent of last five years of fuel” (measured in \$USD) vs “use of fuel in the city” (measured in miles per gallon). Considered instances: BMW, Chrysler Group LLC, FCA US LLC, General Motors, Mazda, Mercedes-Benz, Nissan, Rolls-Royce, Toyota, and Volvo.....	146
Figure 5-12 Scatter plot for the correct instances using the decision trees classifiers of variable “spent of last five years of fuel” (measured in \$USD) vs “use of fuel in the city” (measured in miles per gallon). Considered instances: Ford Motor Company, Maserati, Mitsubishi Motors Co, Porsche, Subaru, and Volkswagen Group.	147
Figure 5-13 Scatter plot for the correct instances using the decision trees classifiers of variable “spent of last five years of fuel” (measured in \$USD) vs “use of fuel in the city” (measured in miles per gallon). Considered instances: Ferrari, Honda, McLaren Automotive, Pagani Automobili S, Quantum Fuel System, Roush, Subaru, Volkswagen, and Aston Martin.	147
Figure 5-14 Scatter plot for the incorrect instances using the decision trees classifiers of variable “spent of last five years of fuel” (measured in \$USD) vs “use of fuel in the city” (measured in miles per gallon).....	148
Figure 5-15 Scatter plot for the correct instances using the decision trees classifiers of variable “spent of last five years of fuel” (measured in \$USD) vs “annual cost of conventional fuel” (measured in \$USD). Considered instances: BMW, Chrysler Group LLC, FCA US LLC, General Motors, Mazda, Mercedes-Benz, Nissan, and Toyota. ...	148
Figure 5-16 Scatter plot for the correct instances using the decision trees classifiers of variable “spent of last five years of fuel” (measured in \$USD) vs “annual cost of conventional fuel” (measured in \$USD). Considered instances: Ford Motor, KIA, Maserati, Mitsubishi Motors Co, Subaru, and Volkswagen Group.	149
Figure 5-17 Scatter plot for the incorrect instances using the decision trees classifiers of variable “spent of last five years of fuel” (measured in \$USD) vs “annual cost” (measured in \$USD). Considered instances: Audi, General Motors, Maserati, Volkswagen Group, and Aston Martin.	150
Figure 5-18 Scatter plot for the incorrect instances using the SVM classifiers of variable “spent of last five years of fuel” (measured in \$USD) vs “the use of fuel in the city” (measured in miles per gallon).....	150
Figure 5-19 Scatter plot for the correct instances using the SVM classifiers of variable “spent of last five years of fuel” (measured in \$USD) vs “the use of fuel in the city” (measured in miles per gallon).	151
Figure 5-20 Confusion matrix for decision tree classifier showing true class vs. predictive class.	152
Figure 5-21 Confusion matrix for SVM classifier showing true class vs. predictive class.	152
Figure 5-22 Confusion matrix for bagged decision trees classifier showing true class vs predictive class.	153
Figure 5-23 Standardized values used for the parallel coordinates plot of categorical instances of the fuel economy data for selection of features.	155

Figure 5-24 Normalized values used for the parallel coordinates plot of categorical instances of the fuel economy data for selection of features.	155
Figure 5-25 Membership function plots for the fuzzy c-means clustering. From top to bottom: cluster 1, 2, and 3.	157
Figure 5-26 Fuzzy c-means partition of 3 clusters plot.	157
Figure 5-27 Scatter plot for the correct instances using ensemble bagged trees classifier of variable “recommended customer price” (measured in \$USD) vs “processor number” (unit number). All classes included.	162
Figure 5-28 Scatter plot for the incorrect instances using ensemble bagged tree classifier of variable “recommended customer price” (measured in \$USD) vs “processor number” (unit number). Considered instances: Intel Celeron® Processor 1000 Series, Legacy Intel Core Processors, Legacy Intel® Pentium® Processor, and Legacy Intel® Xeon® Processors.	162
Figure 5-29 Scatter plot for the correct instances using ensemble bagged trees classifiers of variable “recommended customer price” (measured in \$USD) vs “number of cores” (unit). Considered instances: Intel® Atom Processor C Series, Intel Itanium® Processor 9100 Series, Intel® Xeon Phi x200 Product Family, Intel® Xeon® Processor D Family, Intel® Xeon® Processor E3 v3 Family, Intel® Xeon® Processor E5 Family, Intel® Xeon® Processor E5 v2 Family, Intel® Xeon® Processor E5 v3 Family, Intel® Xeon® Processor E5 v4 Family, Intel® Xeon® Processor E7 Family, Intel® Xeon® Processor E7 v2 Family, Intel® Xeon® Processor E7 v3 Family, Intel® Xeon® Processor E7 v4 Family, Intel® Xeon® Processor W Family, Intel® Xeon® Scalable Processors, Legacy Intel® Celeron® Processor, Legacy Intel® Core Processors, Legacy Intel® Pentium® Processor, and Legacy Intel® Xeon® Processors.	163
Figure 5-30 Scatter plot for the incorrect instances using ensemble bagged trees classifiers of variable “recommended customer price” (measured in \$USD) vs “number of cores” (unit). Considered instances: 5th Generation Intel® Core i5 Processors, 7th Generation Intel® Core i3 Processors, 7th Generation Intel® Core i3 Processors, Legacy Intel® Celeron® Processor, Legacy Intel® Core Processors, Legacy Intel® Pentium® Processor, and Legacy Intel® Xeon® Processors.	164
Figure 5-31 Scatter plot for the correct instances using ensemble bagged trees classifiers of variable “recommended customer price” (measured in \$USD) vs “temperature” (C°). All classes included.	164
Figure 5-32 Scatter plot for the incorrect instances using ensemble bagged trees classifiers of variable “recommended customer price” (measured in \$USD) vs “temperature” (C°). Considered instances: 4th Generation Intel® Core i5 Processors, Intel® Xeon® Processor E3 v3 Family, Intel® Xeon® Processor E5 v2 Family, Legacy Intel® Core Processors, and Legacy Intel® Xeon® Processors.	165
Figure 5-33 Confusion matrix for the ensemble bagged tree classifier showing true class vs. predictive class.	166
Figure 5-34 Standardized values used for the parallel coordinates plot of categorical instances of the CPUs data for selection of features.	168

Figure 5-35 Normalized values used for the parallel coordinates plot of categorical instances of the CPUs data for selection of features..... 169

Figure 5-36 ROC curve plot showing the misclassification of the observation Intel® Celeron® Processor J Series. 170

Figure 5-37 Population growth using GA for feature selection..... 171

Figure 5-38 Clusters found for the CPUs dataset. Upper (a), lower (b). 173

Figure 5-39 Surface plot for coefficient determination of predictive significant values..... 174

Preface

This PhD thesis has been written following the requirements established by the University of Glasgow and the corresponding programme inside the School of Engineering.

As the world moves further into the digital age, technological advancements grow, and manufacturing products become challenging. There will be a greater need to develop methods and gain a solution to high-impact problems, like achieving self-aware, self-adjust, and self-optimize features to a manufacturing process. This work is a first step in this direction.

This research work stemmed from the collaboration and supervision of Professor Yun Li and his research group, whom has been working over the last years in developing ways of applying computational intelligence to address agile manufacturing. My main contribution to his work has been on improving the process of analysing the data to predict potential customer needs and wants to be used as inputs to customizing product designs agilely.

Acknowledgements

Foremost primarily, I am forever in debt and thankful to Professor Yun Li for his supervision, support, experience, patience, and knowledge that without all these the work would have never been possible.

Secondly, I would like to thank the Mexican Council of Science and Technology (CONACyT) for sponsoring in full my PhD studies. Thank you so much for supporting me and hundreds of Mexican students every year.

I would like to express my gratitude and appreciation to Dr Cindy Goh, Dr Leo Yi Chen, Professor Hongnian Yu, Professor Xifan Yao, Professor Wei-neng Chen, Dr Lin Li, Dr Ying Liu, Joo Hock Ang, and Wuqiao Luo for their support and contribution through all the different stages of my research. Collaborating with each one of them has been an honour and this work reflects their support, knowledge, and experience.

Special thanks to my parents Noma and Alfredo, for their constant support and advice on decisive moments.

Most of all I am forever grateful with my beloved wife Hari Datta (Cristy) for being there all the time supporting, listening, loving, caring, laughing, and guiding. Words can never express how grateful I am for all the sacrifices you have made these 4 years and how happy I am for being part of your life and family.

To the family and friends I made in Glasgow, I am the luckiest man for having the opportunity of being part of their lives. Here I include the Krishna family living in Lesmahagow, Anna, Jose, Laura, Andrea Pizzone, Andrea Benecchi, Dan, Marilena, Jose Luis, Raghunath, Prana, Peter, Lisa, Eughan, Shaun, Luis Salinas, and if I miss someone do not worry, memories are engraved in my soul.

I dedicate this to my beautiful daughter Hari Prema, to Walther Enrique and David. Pursue anything that makes you happy; the greatest achievements can be done with

dedication, hard work, and being constant. I would like to dedicate this work as well to my brothers Alexis, Alison, Randy, and Benji as an example that if you dream big the satisfaction is bigger, and mostly creativity can play a key role. Hare Krishna!

Author's Declaration

I declare that this thesis work named “Predicting Potential Customer Needs and Wants for Agile Design and Manufacture in an Industry 4.0 Environment” has been composed solely by myself and that it has not been submitted, as a whole or in part, in any previous application for a degree. The work contained herein is my own except where explicitly stated otherwise in the text.

Parts of this work have been published in:

Paper Title	Paper Type	Authors	Year/Name of Proceeding
Industry 4.0 with cyber-physical integration: A design and manufacture perspective	Conference	Alfredo Alan Flores Saldivar, Yun Li, Wei-neng Chen, Zhi-hui Zhan, Jun Zhang, Leo Yi Chen	2015 21st International Conference on Automation and Computing (ICAC)
Self-organizing tool for smart design with predictive customer needs and wants to realize Industry 4.0	Conference	Alfredo Alan Flores Saldivar, Cindy Goh, Wei-neng Chen, Yun Li	2016 IEEE Congress on Evolutionary Computation (CEC)
Identifying smart design attributes for Industry 4.0 customization using a clustering Genetic Algorithm	Conference	Alfredo Alan Flores Saldivar, Cindy Goh, Yun Li, Yi Chen, Hongnian Yu	2016 22nd International Conference on Automation and Computing (ICAC)
Attribute identification and predictive customisation using fuzzy clustering and genetic search for Industry 4.0 environments	Conference	Alfredo Alan Flores Saldivar, Cindy Goh, Yun Li, Hongnian Yu, Yi Chen	2016 10th International Conference on Software, Knowledge, Information Management & Applications (SKIMA)https://doi.org/10.1109/SKIMA.2016.7916201

Energy-Efficient Through-Life Smart Design, Manufacturing and Operation of Ships in an Industry 4.0 Environment	Journal	Joo Hock Ang, Cindy Goh, Alfredo Alan Flores Saldivar, Yun Li	Energies, Volume 10, Issue 5 (May 2017). Impact Factor 2.676
---	---------	---	--

I confirm that appropriate credit has been given within this thesis where reference has been made to the work of others.

Alfredo Alan Flores Saldivar

Chapter 1 Introduction

In the digital age enabled by information and communications technology and the Internet, the manufacturing sector has been exposed to various circumstances that are ever more significantly impacted upon by customer needs and wants, the inclusion of advanced digital technologies allow innovation to improve and individualise the customer experience by meeting these needs and wants [1]. These circumstances have led companies to react with a strong customer focus, short-cycle adoption, and batch-sizes reduction [2]. The Internet is changing the production floor with more paradigms leading to advancements in how products are designed, customised and manufactured. Present technologies, such as the Internet-of-things (IoT), cyber-physical systems (CPS), cloud-based manufacture, Internet of services (IoS), big data, and smart manufacturing, are driving the advent of the “Fourth Industrial Revolution”, i.e., “Industry 4.0” (i4) or Industrie 4.0 as coined in German [3].

Design and manufacture, as well as service and engineering management, strategies that rely on only the manufacturer’s own decisions without considering the customer’s individual needs, are experiencing challenges attracting the customer’s wants in the Internet era. In this ever more connected society, individualized products and services become more in demand than mass-produced ones [4]. Taking this trend into account, manufacturers are considering customer satisfaction by focusing on design conception and flexible production [5]. This is one of the major principles of i4, where designs are obtained beforehand with the power of internet-based designs, data mining, collaborative systems, and CPS. Agile design and manufacture are considered part of flexible digital manufacturing, where customer-oriented production and knowledge-driven technologies enable agile mass customization, these can be compared with a mass-production when trying to save time and costs [6].

These developments and trends lead to the investigation of what Industry 4.0 will impact on the ways products are designed and manufactured for achieving mass customization. This chapter of the thesis will first discuss the importance of this “industrial revolution”, and a way forward with i4 concepts and approaches. Gaps between current manufacturing systems will also be discussed, together with challenges achieving mass customization in an i4 environment, hence identifying the research problem to be tackled in the work presented in this thesis. The aims of this research and contributions are then outlined.

1.1 Industry 4.0

The first three industrial revolutions came about as a result of centralization for production. Now, businesses are investigating global networks that incorporate their machinery, warehousing systems and production facilities in the shape of a cyber-physical system, comprising “smart machines”, storage systems and production facilities capable of autonomously exchanging information, triggering actions and controlling each other independently [7]. These technologies form a “smart factory” that would allow individual customer requirements to be met, whilst efficiency obtained in automated production is maintained. This means that even one-off items or a product of a batch size of one can be manufactured profitably.

Different from what other smart technologies, digitalization, and future manufacturing perspectives might propose, some of the relevant aspects of i4 are described in the bullet points below, according to [8].

- **Innovative economy.** The key aspect in the way businesses are conducted in the digital era are leading to efficient ways of exchanging information, and most of all decision making. This is owing to upgraded value and supply chains with efficient information flows, which will be discussed in detail later in this chapter.

- **Solution to current challenges for manufacturers.** Industry 4.0 perspective gives the opportunity for companies to adapt to the ever-changing global market and be more responsive to business trends and societal demands. Here is also included the complexity of manufacturing products, and shorter product life cycles, and the use of data to the production floor turned innovation floor for producing a more informed product and helping with the decision-making process.
- **Customer-centred production.** Individualized production based on single users' demand is a key feature of smart technologies. Digitalization is driving customization, allowing faster design processing and alterations for meeting changing customer needs and wants.
- **Human-centred production.** In i4 vision, humans play a centre role, despite what the technological revolution implies a complete substitution of human-labour by the extensive use of machines. Industry 4.0 stipulates only to minimize manual tasks that can be done faster and simpler by machines, but workers will participate in supervision what machines are doing, which means that interaction between humans and machines is essential under i4 principles.

Summarizing the above relevant aspects, the key characteristics i4 brings to the current state of manufacturing are decision-making processes becoming smart, adaptive businesses models, customization, and human-interactive digital systems. In this way, customer-centred and human-centred production are differentiated because of the context of customization as a driver for i4, human-centred production here means that working people inside the manufacturing processes will play a key role, not as customers, but as providers of intellect, expertise, amongst other valuable tasks.

1.1.1 Why Industry 4.0 Is Important

The first three industrial revolutions came about as a result of mechanization, electricity and information technology. Now, with the digital flexibility and Internet connectivity, the introduction of the Internet of Things and Services into the manufacturing environment is ushering in a “Fourth Industrial Revolution”, or i4 for short. This is the first “industrial revolution” that is engineered before it takes place, promising that with it businesses will establish global networks that incorporate their machinery, warehousing systems and production facilities in the shape of a cyber-physical system. In a manufacturing environment, the CPS comprises smart machines, storage systems and production facilities capable of autonomously exchanging information, triggering actions and controlling each other independently.

Such a “smart factory” will allow individual customer requirements to be met, whilst efficiency obtained in automated production is maintained, meaning that even one-off items or products or components of a batch size of one can be manufactured profitably. In i4, dynamic business and engineering processes would enable last-minute changes to production and offer the ability also to respond flexibly to disruptions and failures. End-to-end transparency is provided over the manufacturing process, also facilitating optimized design and decision-making.

Despite that manufacturing companies generally oppose to growing global competition, more individualized customer demands, new technologies and rapid technological progress, as well as strict environmental regulations, i4 will dynamically enable business and engineering processes to deal with last-minute requirements or changes to production and deliver the ability to respond flexibly to disruptions and failures. These trends lead to an increase in product variety, shorter product life cycles, uncertain and fluctuating demands, as well as higher cost pressure. Figure 1-1 illustrates how mass production to mass customization is likely to shift in future times [9].

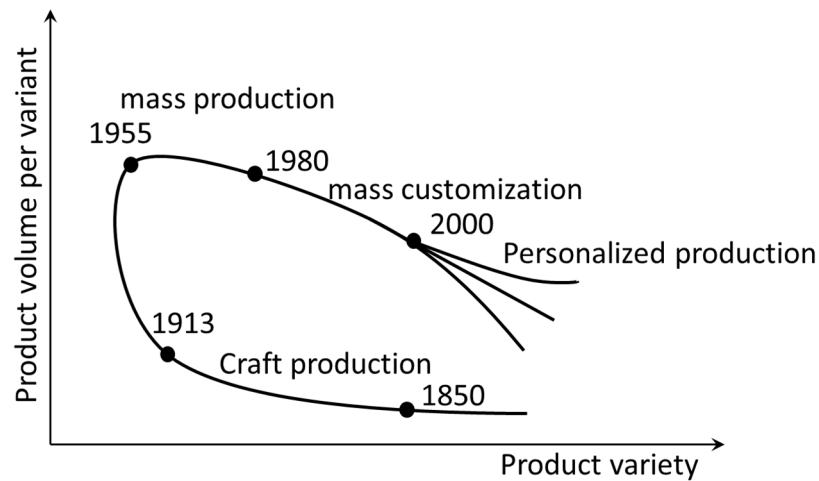


Figure 1-1 Trend to mass customization according to [9].

Moving forward, i4 will lead to new ways of creating value and novel business models. For example, it will provide start-ups and small businesses with the opportunity to develop and provide downstream services. To economies developed and developing, i4 will reduce factory-floor requirements and help progress humanity.

1.1.2 Components of Industry 4.0

What today are named “industrial revolutions” started with the incorporation of manufacture. Technological advances have carried paradigm shifts ever since. Figure 1-2 shows those advances [3].

and services by the internet (iv). Via decentralising intelligence, object networking and independent process management interact with the virtual and real worlds, heralding a crucial new aspect of the future industrial production process that integrates the above four processes. In short, i4 represents a paradigm shift from “centralised” to “decentralised” production, a reversal of the logic of production process thus far. The design principles of i4 components are shown in Table 1–1[11].

Table 1–1 Design principles of each Industry 4.0 component.

	Design & customisation	CPS	Smart Factory	IoT	IoS
Modularity	X	-	-	-	X
Interoperability	X	X	X	X	X
Real-Time Capability	?	-	X	-	-
Virtualisation	X	X	X	-	-
Decentralisation	X	X	X	-	-
Service Orientation	X	-	-	-	X

For each design principle is necessary to describe how it matches with i4 components:

- **Modularity:** modular systems can flexibly adapt to changing requirements by replacing or expanding individual modules. For that reason, modular systems can be easily adjusted in case of seasonal fluctuations or changed product characteristics. Another concept for Smart Factory plant is the Plug&Play principle, which can also add new modules. Via the IoS, new modules are identified automatically and can be utilized immediately, based on standardized software and hardware interfaces [12].
- **Interoperability:** an important enabler of i4, because, for companies running with i4 principles, CPS and humans are connected over the IoT and IoS. A success factor for communication will be standards, between CPS of various manufacturers. In the context of Smart Factory plant, interoperability means that all CPS within the plant (work-piece carriers, assembly station and products) are able to communicate with each other “through open nets and

semantic descriptions”, for design and customization is of importance because here is where the virtual part of the product is linked and feedback to the process in order to reach individual necessities for customers, therefore customizing it [12].

- **Real-Time Capability:** for organizational tasks, it is necessary that data is collected and analysed in real time. In the Smart Factory, the status of the plant is permanently tracked and analysed. Thus, the plant can react to the failure of a machine and reroute products to another machine. Yet for design & customisation, it’s still debated if can be processed real-time, or if it’s suitable for the physical process [13].
- **Virtualization:** this means that CPS are able to monitor physical processes. Data is collected from the sensors allocated in various parts of the physical process, then this sensor data is linked to virtual plant models and simulation models. Thus, a virtual copy of the physical world is created. In the Smart Factory plant, the virtual model includes the condition of all CPS. In case of failure, a human can be notified. In addition, all necessary information, like next working steps or safety arrangements, are provided. For design and customisation virtualization means that once created the virtual copy of the product, here it can be modified with different settings already fed from the customer needs and wants analysis through big data. Hence, humans are supported in handling the rising technical complexity [12].
- **Decentralization:** rising demand for individual products makes it increasingly difficult to control systems centrally. Embedded computers enable CPS to make decisions on their own. Only in cases of failure tasks are delegated to a higher level. For quality assurance and traceability, it is necessary to keep track of the whole system at any time. In the context of Smart Factory plant,

decentralization can be exemplified as the Radio Frequency Identifier (RFID) tags “tell” machines which working steps are necessary. Therefore, central planning and controlling are no longer needed. For design & customisation means that, based on the selected modifications to reach customer needs and wants, the decision within the system or process enables the product to be manufactured [13].

- **Service Orientation:** services of companies, humans, and CPS are available over the IoS and can be utilized by other participants. Smart Factory plant is based on a service-oriented architecture in which service can be offered internally and across company borders. All CPS offer their functionalities as an encapsulated web service. This result on the product-specific process operation, that can be composed based on the customer specific requirements provided by the RFID tag making more reliable the process of designing and therefore customizing products [13].

Based on technological concepts where Design & Customisation, IoS, IoT and CPS come together and facilitates the vision of what a Smart Factory is, and as discussed above the key is the decentralized system, which within the modular structured Smart Factories, the CPS monitor physical processes, create a virtual copy of the physical world and make decisions. The connection and communication between the CPS and the IoT allow co-operating with each other and humans in real time. Internal and cross-organizational services are offered and utilized by participants of the value chain via the IoS.

Several industries in Germany show interest in developing and lead a well-integrated methodology to optimize connection through the Internet and smart devices pursuing a service-oriented strategy and strong customization of products under the conditions of high flexible production [3]. With the introduction of methods that can be

adaptable, self-learning, self-aware, self-predicted, self-optimized, self-configuration and self-maintained, allow the required automation technology to be improved, which outstands as an innovative feature for business models that totally changes the way of making products and services [14].

Once defined what i4 is and its components and design principles, the next section discuss briefly the Smart Factory concept.

1.1.3 Smart Factory

Research and developments are heading the smart industry to a well-structured model which can be optimized and automated. Smart factory products, resources and processes are characterized by the CPS, providing significant real-time quality, time, resource, and cost advantages in comparison with classic production systems. The smart factory can be designed according to sustainable and service-oriented business practices, for which those rely upon adaptability, flexibility, self-adaptability and learning characteristics, fault tolerance, and risk management.

High levels of automation come as standard in the smart factory: this being made possible by a flexible network of CPS-based production systems which, to a large extent, automatically oversee production processes. Flexible production systems which are able to respond in almost real-time conditions allow in-house production processes to be radically optimized. Production advantages are not limited solely to one-off production conditions but can also be optimized according to a global network of adaptive and self-organizing production units belonging to more than one operator.

Smart factory production brings with it numerous advantages over conventional manufacture and production. These include:

- CPS-optimized production processes: smart factory “units” are able to determine and identify their field(s) of activity, configuration options and production conditions as well as communicate independently and wirelessly with other units.
- Optimized individual customer product manufacturing via an intelligent compilation of ideal production system which factors account product properties, costs, logistics, security, reliability, time, and sustainability considerations.
- Resource efficient production.
- Tailored adjustments to the human workforce, so that the machine adapts to the human work cycle.

Conversely, despite the significant penetration of cloud computing and smart manufacturing approaches, many companies are staying out of it, the reason seems to be the resistance of users because sometimes the low time response of some applications. Depending on the task users have, delays on data transferring or applications may affect the interaction between the system and end-user, so the biggest challenge cloud computing is facing at present is having a faster link to load and download information [15].

1.2 Aims of This Research

In this work, the main focus for considering smart technologies and i4 principles for manufacturing is to develop a methodology capable of addressing customization under smart manufacturing principles. This needs to go upstream in the value chain, notably, to the product design stage, for example. Highlighted in the previous section of this chapter, the role of data on the use of the manufactured goods have been

underlined as a key aspect of the digital revolution [16] [17]. In this sense, such downstream data is fed back to the upstream and can thus be considered as the starting point of how to connect information to customized product designs for smart manufacturing.

This idea is used for establishing a scope since i4 and smart manufacturing comprise a vast number of challenges and work to be carried out, as the i4 concept is still under development. Moving forward with the above-presented challenges for mass customization, we anticipate that an effective integration of concepts, cutting-edge technologies capable of responding to complex processes, and simple or intuitive ways of making design and manufacture more smartly have been the missing gaps. Thus, while it has been recognised that CPS and IoT are considered to be the main drivers of the fourth industrial revolution, data are considered to be the driver of customization since data analytics can lead to meet individual needs and wants through virtual product designs.

In this sense, for closing the value chain loop from design to manufacturing and to IoT-based services, it is desirable to select the best product attributes for the design in anticipation, meaning that the selection of the best product design that matches individual needs is chosen with prediction. This thesis, therefore, aims at improving ways of analysing the data for an informed representation of customer needs and wants on a manufacturing system, such that this helps the decision-making process of selecting the best product design for manufacture. This is the main reason why this work is focused on data analysis with artificial intelligence (AI), but also concentrates on smart environments that match i4 principles. In practice, machine learning approaches can be used to obtain meaningful and useful information about customers' behaviour, needs, and wants. With this information, then several aspects of design elements can be obtained directly from the data analysis using AI. Once the results

are extracted, a decision for i4-ready design and manufacture can be made with the collected information and experts' opinion.

Specifically, this thesis will address one of the customization issues in smart design: prediction of customer needs and wants for smart production. Thus, the main objectives of this thesis are:

1. Develop an AI-based methodology to automatically predict the design attributes that best reflect what customers need and want in a product for customised manufacture;
2. Obtain a model capable of accurately predict customer needs and wants for at least 85% of classified design attributes;
3. Contribute by identifying effective ways of achieving customization for i4 and smart manufacturing;
4. Develop a machine learning approach that would explain at least 85% of the variance when building a model to predict customers' needs and wants;
5. Obtain an analysis capable of determining the best design attributes/features that can be utilized to predict customer needs and wants;
6. Contribute useful knowledge for a closed-loop value chain to advise individualized production in smart manufacturing and i4 environments.

These objectives are for efforts on closing the gap between smart design and manufacture for i4 and its commercial potential. The determination of the prediction interval of at least 85% is taken from [18], where is explained that the region where

true outputs of new attributes (in this case) might fall, and the use of this interval give us the opportunity to validate against trained data if predictions are good or not. Moreover, when true classified values fall into this region, is still possible to perform a separation of reliable predictors, and minimize the rate of false positives, this can be obtained with the adjustment of the model.

To begin, a critical review will be carried out to attempt some answers to the following questions:

1. Where in the industry value chain most value is added?
2. What are the major benefits of predicting needs and wants in i4 environment to the customer?
3. How to design smart products agilely in this value chain?
4. What benefits will predict customer needs and wants in i4 environment bring to the manufacturer?
5. What are major challenges to predict customer needs and wants in i4 environment?
6. How will i4 add most value and/or efficiency?

With the above questions and objectives, this thesis mainly contributes to improving the process of analysing the data to predict potential customer needs and wants to be used as inputs to customizing product designs agilely.

This thesis aims at agile manufacturing, which is an approach to manufacturing to focus on meeting the needs of customers while designing and maintaining with high

standards of quality and controlling the overall of production. The analysis in the thesis is therefore focused on the reduction of the number of design attributes selected in a predictive way through a closed-loop framework that integrates many of the key drivers of Industry 4.0 and smart manufacturing principles (IoT, cloud computing, CPS, data analytics, digital aided design, etc.). It also integrates concepts like computer automated design (CAutoD) and AI to help improve the decision-making process of customizing products according to subconscious individual requirements. The motivation on the used case studies or datasets for performing the analysis lies on the concept that will be discussed in detail in Chapter 2, section 4, i.e. big data and business informatics. Most of the datasets used to perform data analytic tools come from sales/markets environments because of the nature of the problem presented and predicting what customer needs and wants are. In this context, the collected data can lead us to obtain valuable information about individual needs, and turn such needs and wants into design attributes for customizing products.

1.3 Outline of the Thesis

The remaining chapters are organized as follows: Chapter 2 gives an overview of literature and a critical review of research in the area of this work, where the review includes smart manufacturing developments, Cyber-Physical Integration realising smart products, big data and business informatics for i4, AI for smart manufacturing, and finalizing with a summary and study cases. Chapter 3 presents the methods used for predicting attributes under smart design principles, here the Cyber-Physical Integration, considered Machine Learning approaches, Smart Design under i4 principles, and a summary and motivation are covered. Chapter 4 includes the proposed frameworks for predicting potential needs and wants, the different aspects and improvements are presented as sections in this chapter, which are value chain for predicting potential customer needs and wants, AI closed-loop, Classification learner, Genetic search, and summary. In Chapter 5, the application, evaluation of

machine learning approaches, and case studies are presented, where each dataset is introduced, a motivation of selecting the case studies is given, the data analysis and results for each dataset are shown, and the obtained results are summarized in the last section of this chapter. Finally, in Chapter 6, the conclusion and future work are discussed.

Chapter 2 Literature Review

2.1 Smart Manufacturing Developments

It is stated in [19] that smart manufacturing represents a collection of technologies that promote strategic innovation having an impact on the existing manufacturing industry by converging technology, humans, and information. The innovation for smart manufacturing has been also spread thanks to the extensive use of internet technologies, allowing faster communication between customers, stakeholders, machines, and shop floor workers; this communication enables actions towards better-informed decisions.

Part of the main goals of i4 is the concept of Smart Factory as the most complete development, in which all the cutting-edge technologies take place as one of the main drivers of the fourth industrial revolution. According to [20] a Smart Factory is identified as a manufacturing solution that provides flexibility and adaptability to production processes, these capabilities give a solution to the encountered problems in a dynamic and faster way where complexity increases and traditional ways of making products are not possible. Automation is essential to maintain production according to desired standards and quality in a Smart Factory, here information and Internet technologies, mechanics, and internet applications can lead to optimize manufacturing resources, resulting in minimizing the waste of resources and unnecessary labour.

Developments and design involved inside Smart Factory concept required a background vision, this vision was first addressed in [21], where it was described a physical world that is connected and interlaced with actuators, sensors, computer elements and displays, and all these elements are seamlessly embedded into daily life objects. A network is the mean of connection between objects, machines, and people, which then this vision was transferred to manufacturing issues. Thanks to the

evolution of information and communications technology, virtual and digital developments, and global network technologies the factories are experiencing a change because of the fusion of physical and virtual worlds allowing smart technologies to drive this paradigm shift [22].

Thanks to smart factory development distribution, real-time collection, and access of manufacturing relevant information can be retrieved and accessed anytime and anywhere [20]. These developments enable decentralized information and communication structures for smart manufacturing since the process can handle faster changes because of the vertical and horizontal integration of information systems, an example for this is the assignment of material and flow of information inside and outside an enterprise. In terms of context-aware according to [20] the applications in a Smart Factory need to answer the following questions:

- 1) Identification stage → how is an object identified?
- 2) Positioning stage → where is an object located in the factory?
- 3) Status knowledge → what is the status or situation of an object?

The above questions lead to consider some challenges that arise with these topics. These challenges are described as follows:

- Identification: information of the real world is assigned to objects that are suitable to be identified, tagged, sensed, and establish a connection to a facility. The object is identified, and a task is assigned to be processed in rough industrial environment.
- Localization: having information about the position of objects (tools, components, materials, etc.) can improve the process and reduce idle times.

For smart environments, this positioning system needs to have a certain level of robustness and work on large scale in accordance with environmental influences, noise of dust, electromagnetic fields, etc.

- Status knowledge: users need to be informed or as discussed before, the context-awareness of objects is key in smart factories to know the status of processed jobs in the system.
- Update of smart management systems: status or location of an object has to be communicated to the systems inside a smart factory periodically.
- Support for different queries: a smart factory has to support different types of queries (object-based, location-based, temporal, and combination of all types) as part of assistance systems.
- Integration of heterogeneous information: different systems inside a company can cause challenges when interfaces, information models, and data formats are not based on a common language in order to achieve synchronization. This challenge can be resolved easily by integrating and building a common platform.
- Real-time characterized reaction: in order to give support to people and machines, the information has to be processed in seconds. For this information and communications technology and database management address this challenge.

These challenges encompass and describe how customers and companies communicate with each other by interacting with objects in common. For the challenges above, in [20] is discussed that customers need to be aware only of the status knowledge and localization stage, and companies should manage the rest of

the challenges. Smart Factory vision at the end of the day makes the job easy for companies to make better decisions, to reduce waste, to increase profits, and most important to satisfy customer needs and wants. The next section discusses mass customization paradigm under i4 principles.

2.2 Industry 4.0 - Mass Customization and the Entire Value Chain

Discussed in [11, 12], it is central to the vision of i4 to address individualised production at mass production costs, where mass production costs represent large quantities of products mass-produced, but in i4 individual needs can be met and still get the benefit of the product cost being mass-produced. Because of the increased influence of Internet and globalization, companies worldwide started to consider a shift on how to conduct business and develop strategies [23], leading to the conclusion of include production plans that satisfy customers' needs and wants but as well considering the benefits of mass production efficiency [24]. Mass customisation in manufacturing supply chain has some implications that concerns material flow and information, this leads to the connection between product types and the decoupling point, affecting customer satisfaction [25].

The manufacturing companies today are facing major challenges providing a high level of product variety at mass production costs. Adequate operational systems and machinery need to carry out manufacturing processes capable of dealing with individualised flexibility but at the same time using resources efficiently and ensuring quality as well [23]. Flexibility and autonomous adjustment can be achieved with the CPS, allowing analysis of individuals for future events without reducing reliability to the workpiece once processed which can be automatically adjusted for individual processes [23]. The quality of the final product can be automatically checked by comparing the end-up product with a target or desired data created on the Computer-Aided Design (CAD) system. In this last process we consider that automation can find

a better application if the virtual design is optimised from an initial stage, taking into account what has been proposed in [26, 27] as CAutoD when customers' needs and wants are detected a priori the quality target can be set and compared in a closed-loop.

The importance of i4 for individualised production is that the recent developments in technologies like digital technology, manufacturing technology, and network technology are integrated to boost design-production-management-service [28]. Companies nowadays are realising that customers are getting involved more and more in the design processes, and that puts them in the position of being no longer considered passive buyers, this concluding in the need to address the social element of customer demands by developing flexible production methods that can meet customers' needs and wants of multiple individuals [28]. Moreover, traditional manufacturing production methods currently cannot meet the social aspect of manufacturing development requirements. Simultaneously, market supply chains and manufacturing enterprises share an information barrier, that in this context according to [3] i4 includes two big subjects: 1) intelligent production and 2) a smart factory. This will allow machine fleets to self-organize, and the supply chain to automatically be coordinated.

Without the context of i4 and smart manufacturing, customisation, as considered from the business perspective, requires the operating network to be dynamic because the purpose at the end of mass customisation is to adapt one-to-one, allowing customers to design their needed products themselves [29]. Some of the advantages of mass customisation include:

- Increased cash flow: payment in advance (minimize receivables), minimize inventories...maximise cash flow.

- Maximised market share by maximising customers' satisfaction and number of clients.
- Reduce cost of inventories and material waste: implementing just-in-time, not produce to stock, and minimize inventory of finished goods (make-to-order).
- Shorten time of responsiveness: flexible manufacturing and organization structure can result in adaption to different demands quickly.
- Ability to supply a whole line of services and goods at bottom prices: the key is to differentiate products to particular demands, resulting in wider companies' product lines and minimal risk of obsolete inventory.

The advantages presented above considering the context of i4 and smart factory focus on the technologies not described above: IoT, IoS, and CPS. Those technologies work as enablers of i4 [30] and bring the concept of make-to-order to a different level of manufacture, in which all the advantages presented before came as a result of the interaction between customers and companies both connected to a common network in which a constant feedback is necessary to facilitate the design process and desired quality. In this way, production happens after the customer place an individual order, and the company knows exactly what to produce, involving which material, process involved, quantities, and quality.

Finally, to complement the revision of what customisation for i4 is, the following approaches to mass customisation are highlighted [29]:

- Adaptive customers: standard product can be bought and customers have the option of modifying those by themselves according to their own needs.

- Collaborative customizers: companies create a dialogue with customers to address their needs and wants and then develop customised outputs to meet those needs. Examples of this approach are Nike, Dell and Levi's which basically in each shop a computer system is provided to measure settings in terms of customers' needs, the information is sent to the shop floor where the company produce a custom-fitted good.
- Transparent customizers: companies provide custom products and the customers do not know that a product has been customised for them. This approach can be found in business like Amazon or Netflix, in which each profile is tracked how each individual uses the service and then start to suggest features that customer might find useful.
- Cosmetic customizers: a standard product is produced but packaged differently for each customer. The examples are chip producers that need to use different packages for each customer, like Lays, other retailers or supermarket brands.

The basic approaches shown above describe what customisation can bring to the business perspective of i4 but also consider the manufacturing part. In the next subsection will be presented the value chain concept in accordance with digital manufacturing.

2.2.1 Entire Value Chain

Value chain for i4 is described by many authors like [31] as a further developmental stage in organisation and management of the entire value chain process involved in the manufacturing industry. Digital manufacturing and design draw attention to innovators, those new digitally-enabled technologies that include advances in production equipment, smart finished products, and data tools and analytics across the value chain. As many companies start adopting this information and

communications technology the boundaries between the real world and virtual world are closing the gap to have a more integrated Cyber-physical Production System (CPPSs). The CPPSs work as online networks of social machines, with mechanical and electronic components the communication with each other, is via the network. Smart machines continually share information about stock levels, problems or faults, and changes in orders or demand levels. CPPSs not only network machines with each other, but they also create a smart network of machines, properties, information and communications technology systems, smart products and individuals across the entire value chain and the full product lifecycle.

Value chain concept was popularized and developed by [32], defined as the amount buyers are willing to pay for what a firm provides, the value chain is the combination of nine generic value-added activities operating with a firm, activities that work together to provide value to customers. Then first, value is a subjective experience that is dependent on context, the more the necessity of something, the most value is added to; second, value occurs when needs are met through the provision of products, resources, or services; and finally, value is an experience and it flows from the person (or institution) that is the recipient of resources, it flows from the customer. These concepts point out what is a key difference between a value chain and a supply chain, they flow in opposite directions. Shown in Figure 2-1 is the order fulfilment value chain as a pictorial of the comparison [33].

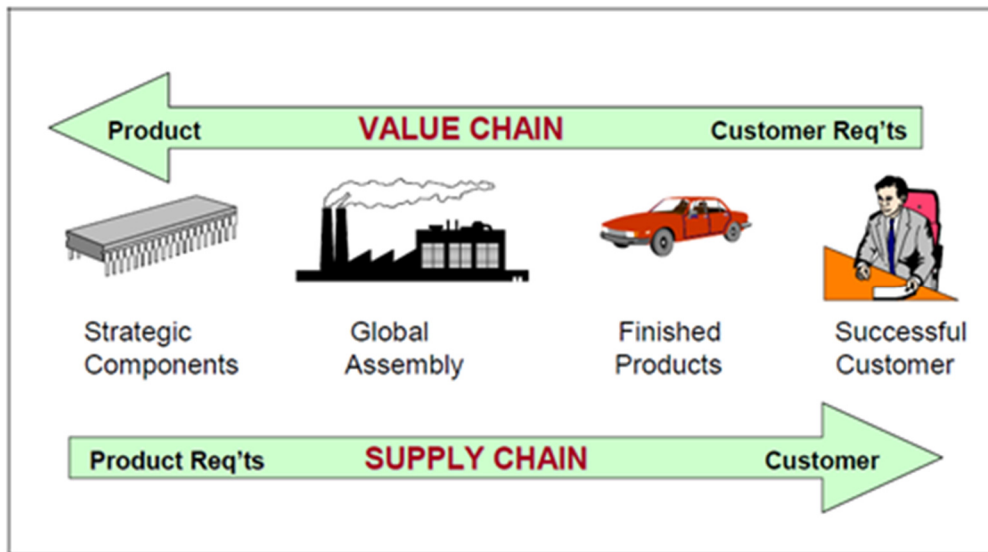


Figure 2-1 A comparison of a Value Chain with a Supply Chain [33].

This framework presented in Figure 2-1 helps to see the loop and constant feedback necessary in customer's needs and wants fulfilment, then for the question stated in the previous chapter as part of aims of this research: Where in the Industry Value Chain, most value is added? Is compulsory to think that customer plays a key role and most precisely that for i4 principles, smarter way of manufacture products adds value, the potential lies on highly customized products at mass production costs. It is expected that i4 allows for a faster response to customer needs than it is possible today. It improves the flexibility, speed, productivity, and quality of the production process. And it lays the foundation for the adoption of new business models, production processes, and other innovations. This will enable a new level of mass customization as more industrial producers invest in i4 technologies to enhance and customize their offerings.

In summary for the above-presented concepts and approaches, it is necessary to discuss what focus smart design can bring when used according to i4 vision. Going through all the revised concepts, technologies and approaches surrounding i4 lead us to include in this revision one of the key technologies and concepts that researchers

[3, 11, 12] discuss about enabling customised designs at larger scale, which is CPS technology. The next section shows a detailed revision of what CPS contributes and means to i4, and some developments will be also discussed.

2.2.2 Gaps Between Current Manufacturing Systems and Industry 4.0

Groover [34] stated that many researchers agreed that manufacturing systems are influenced by different factors, such as the number of workstations, types of operations, system flexibility, and automation level. These factors are used as a baseline to set the fundamentals of i4. The following types of manufacturing systems are included in i4 fundamentals: single-station automated cells, single-station manned cells, automated assembly systems, manual assembly systems, flexible manufacturing systems, and cellular manufacturing systems [34].

- **Single-station automated cells:** These stations or cells are fully automated, and the machines involved are not attended by workers during most of the machine cycles. This type of manufacturing considers production increments and labour costs decrement. The system nonetheless, also targets constant product batches. This type of manufacturing system is the beginning of digitalization on the factory floor and automation but differs with the i4 principles in the lower flexibility for customizing products [35].
- **Flexible manufacturing system:** These are highly automated systems, where several workstations are connected to an automated transport that constantly feeds assembly lines, and the digital part that controls the system is distributed. Workpieces inside this system are identified in the entire production cycle, which enables instant changes in processing. Usage of material, inventories, and maintenance of equipment is improved. Additionally, because of the high flexibility, the system is capable of performing quick responsiveness required to make changeovers. Here, i4

advances this type of system because of the extensive use of computational systems, and the digitalization of the workstations enables workers to bring innovation to the workshop [36].

- **Automated assembly system:** These systems replace human labour with industrial robots (e.g. handling system), to bring full automation of pre-fixed orders and schedule manufacturing of specific products. This system requires high stability without changing product design during the production process. One of the key features here is the massive product demand, which normally handles at least millions, and considered to be more profitable. Similar to assembly systems in i4, components like quality, safety monitoring, and sequence control are automated. Here, i4 shares the automation of quality, control, and mass production, but clearly in a more flexible way.
- **Computer-integrated manufacturing system:** In this system computers control the whole functionality. Ideally, this manufacturing implicates that automation in the factory level involves materials management, design, production line, and distribution. The reduction of error can be detected rapidly with the constant retrieval of information using integrated computers. For i4 principles, this manufacturing system shares the feature of being capable of cooperative automation [37].
- **Reconfigurable manufacturing system:** This system is created for adjusting to sudden changes either in the market or regarding requirements from the same line of products. Six capabilities are identified in these systems: integration ability, modularity, convertibility, customization, diagnosing ability, and scalability. These systems aim to increase the changing response of different requirements, paying more attention to personalized flexibility than production flexibility. These type of systems cope very well with the i4 principles, in the

sense that they seek more personalized or customized features, and bring the flexibility capable of achieving mass customization[38].

Figure 2-2 summarizes these manufacturing systems and compares them with i4 principles. The single-station automated cell is digital and wired to achieve flexibility. It is hard to find the automated assembly system standardized, which is due to the computer-integrated manufacturing system that is executed beforehand. In both reconfigurable and flexible manufacturing systems, customers can order goods based on their ideas. Nonetheless, current flexible manufacturing systems lack real-time responses. It is clear that flexible and reconfigurable manufacturing systems are the closest to what i4 aims to achieve. Hence, the systems depicted in Figure 2-2 are concepts that are difficult to achieve with current manufacturing systems. To realize these manufacturing systems, there needs to be a shift in the way processes are set up. For i4 developments, our research aims to meet some of these concepts in all directions and propose a solution to achieve a process that can be self-configured, self-optimized, self-aware, and help with decision making. This can finally close the gap between current manufacturing systems and i4.

On the other hand, there is still more than one thing to improve on the side of manufacturing, underlining that those improvements had to be the future directions. Many levels need to be scaled up, and the ability to provide consciousness to processes intelligence is extremely difficult.

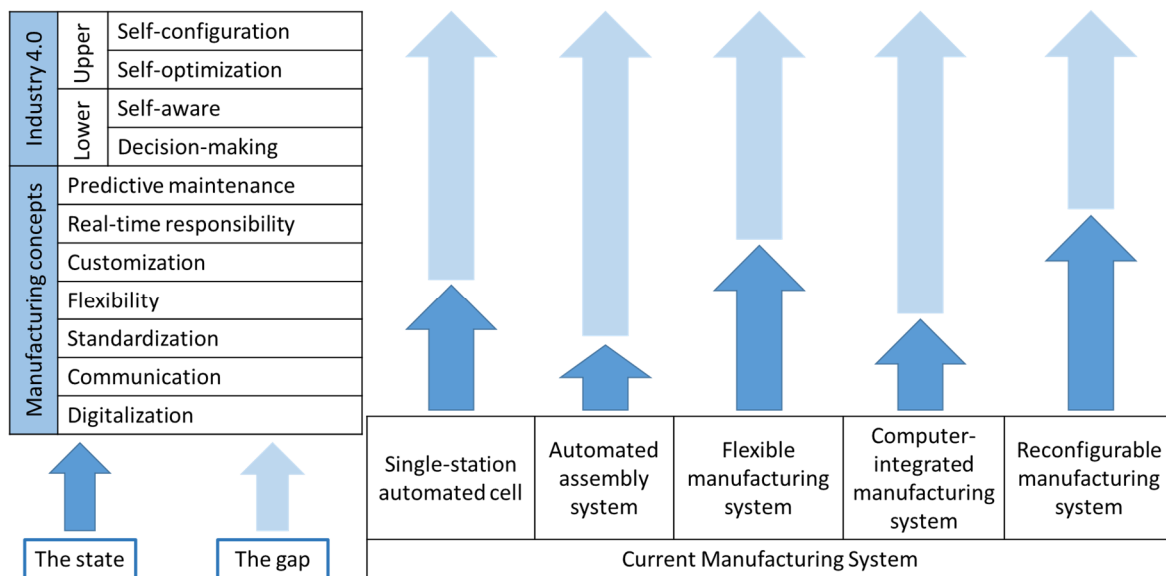


Figure 2-2 Research gap between recent manufacturing systems and i4 [39].

2.3 Cyber-Physical Integration Realising Smart Manufacturing

According to [40], CPS are systems of collaborating computational entities which are in intensive connection with the surrounding physical world and its on-going processes, providing and using, at the same time, data-accessing and data-processing services available on the internet.

CPPS relying on the newest and foreseeable further developments of computer science, information and communication technologies, manufacturing science and technology may lead to the 4th Industrial Revolution, frequently noted as Industry 4.0, which holds a big potential to change every aspect of life.

Concepts like autonomous cars, robotic surgery, intelligent buildings, smart manufacturing and implanted medical devices are just some of the practical examples

that have already emerged as the opportunities that CPS can offer as part of the Research and Developments are leading by several groups [41].

A well-funded approach for Cyber-Physical Integration is shown in Figure 2-3, proposed by [40].

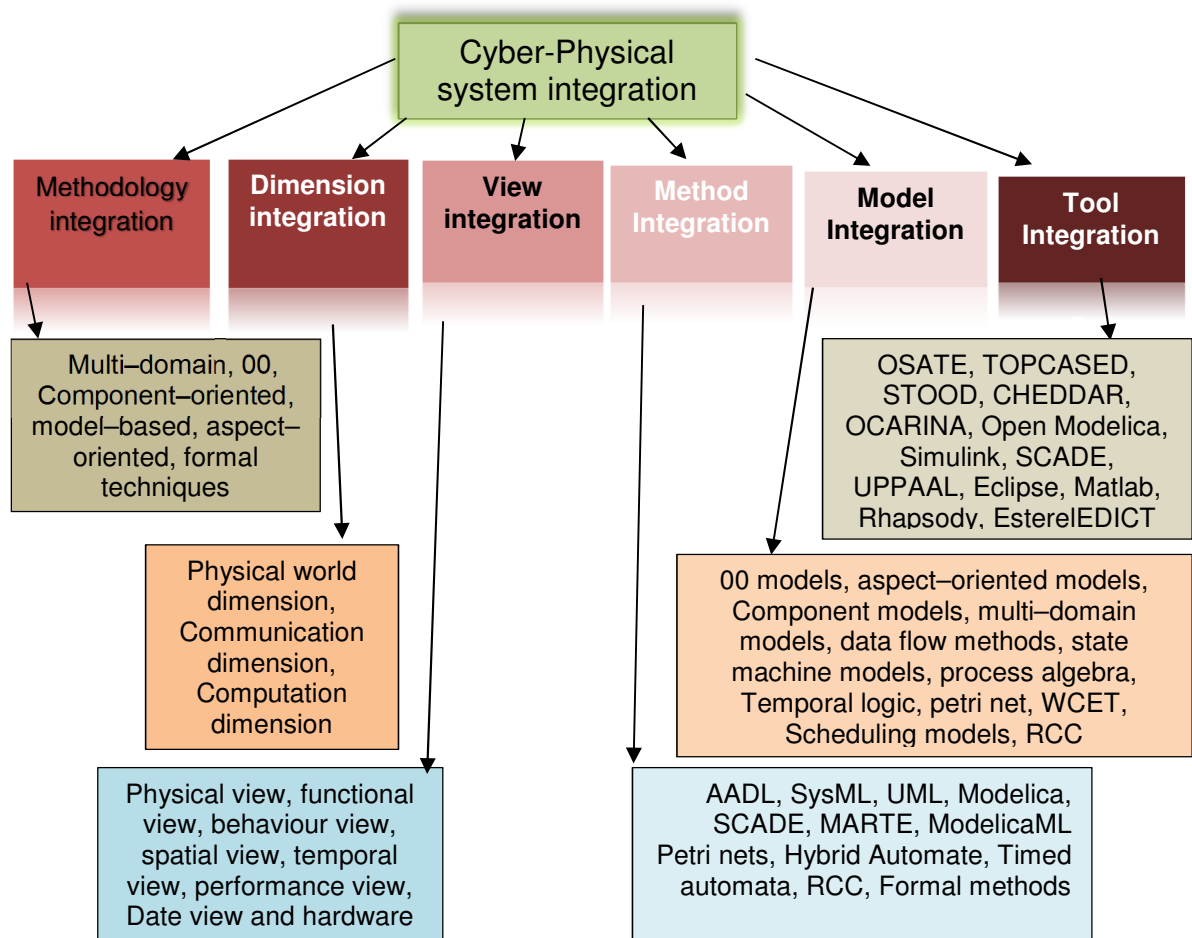


Figure 2-3 Integrated Approach to develop CPS -. [40].

In [42] it is highlighted that analysis is a key issue in current networked cyber-physical system developments, the desire to integrate various objects, design methods and tools, aspect-oriented development methods and tools, multi-domain physical modelling methods and tools, and formal methods that address different aspects of the development process of cyber-physical systems. Cyber-physical systems

specification, modelling and design method integration involves many aspects of integration and different levels:

- The integration of physical world dimension, communication dimension and computation dimension.
- The integrated object-oriented methodology, multi-domain methodology, aspect-oriented methodology and formal techniques.
- The integration of different design views. Views refer here to dimensions used as starting point for modeling and design.
- The integration of the methods used to specify and implement systems requirements.
- The integration of tools that support these methods.
- The integration of physical components and cyber components.
- The integration of different representations.
- The integration of the multiple specification fragments produced by applying these methods and tools.
- Integration between informal specification methods and formal specification methods is desired.

In the following subsection are considered the Model Integration, Methodology and Tool Integration developments as part of the CPS applications. These developments

at the end will address cases of study that will help to develop a specific case to focus on.

2.3.1 Cyber-Physical Integration

CPS is an important component of i4, from the point on which the fusion of the physical and the virtual world comes together. This fusion is a reality with CPS. CPS are “integrations of computation and physical processes. Embedded computers and networks monitor and control the physical processes, usually with feedback loops where physical processes affect computations and vice versa” [43].

Three phases characterize CPS developments [44]:

First generation → includes identification technologies like RFID tags, allowing unique identification.

Second generation → CPS equipped with sensors and actuators with a limited range of functions.

Third generation → able to store and analyse data, equipped with multiple sensors and actuators, also network compatible.

Development: Model Integration

As discussed before, the key in CPS is to develop methodologies that integrate models, techniques, tools that can be used in a design customized within its models and components. Components and Models in cyber-physical systems are heterogeneous, span multiple domains (physical - thermal, mechanical, electrical, fluid..., and cyber-software, computing platforms), and require multiple models to soundly represent physical aspect, the requirements, architectures, behaviour, spatiotemporal constraints , and interfaces, at multiple levels of abstractions [42].

In [45] a new integration model for the OpenMETA suite is proposed, basic design flow is implemented as a multi-model composition/synthesis process that incrementally shapes and refines the design space using formal, manipulated models. Include analysis and testing steps to validate and verify requirements and guide the design process to achieve least complex, therefore the least risky and least expensive solutions. Figure 2-4 shows the proposed design flow by [45].

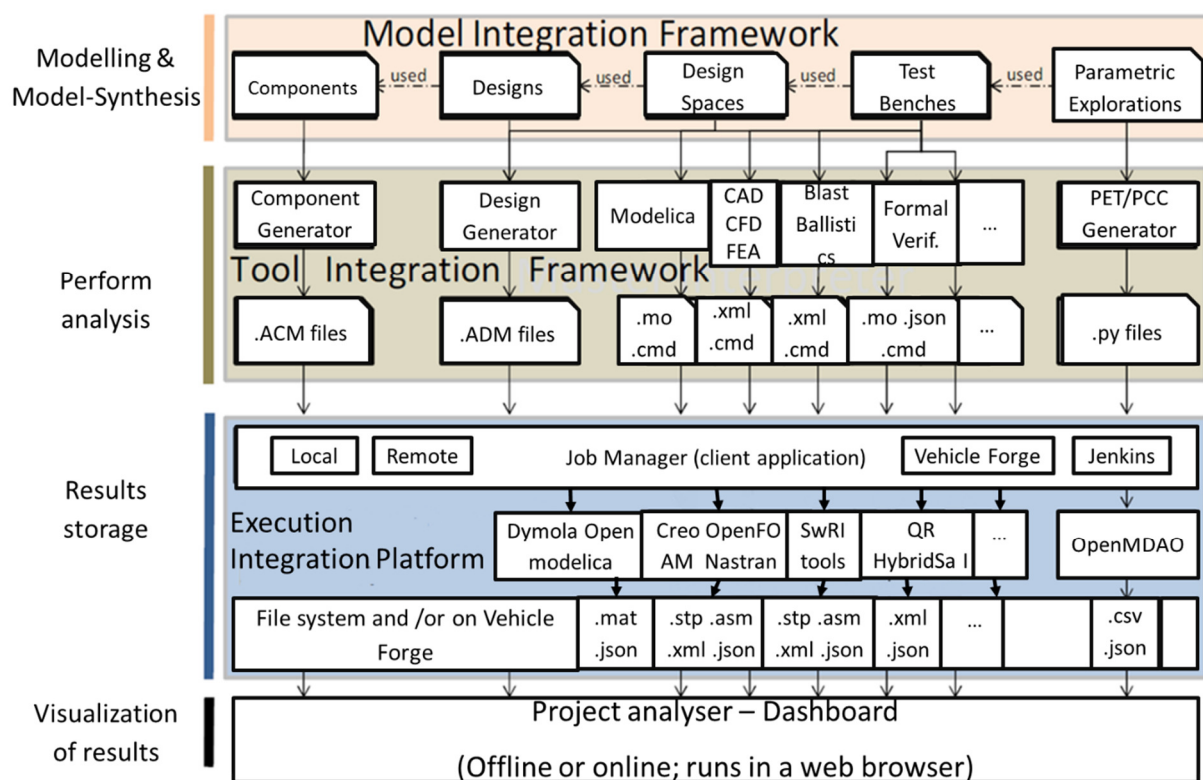


Figure 2-4 Model Integration: OpenMETA framework. - [45].

The main procedures of this design flow sketch the following phases:

- 1) Combinatorial design space exploration using static finite domain constraints and architecture evaluation.

- 2) Behavioural design space exploration by progressively deepening from qualitative discrete behaviours to precisely formulated relational abstractions and to quantitative multi-physics, lumped parameter hybrid dynamic models using both deterministic and probabilistic approaches.
- 3) Geometric/Structural Design Space Exploration coupled with physics-based nonlinear finite element analysis of thermal, mechanical and mobility properties.
- 4) Cyber design space exploration (both HW and SW) integrated with system dynamics.

Development: Method Integration

Many researchers agreed on having a methodology which integrates modelling languages, in order to control Cyber and Physical environments, mathematical models in this sense can bring together those abstractions that are imported from the individual languages and required for modelling cross-domain interactions. Proposed by [45] the language called CyPhyML is constructed as a light-weight, evolvable, composable integration language that is frequently updated and morphed. While these DSMLs may be individually quite complex (Modelica, Simulink, SystemC, etc...) CyPhyML is relatively simple and easily evolvable. This “semantic interface” between CyPhyML and the domain-specific modelling languages (DSML) shown in Figure 2-5 is formally defined, evolved as needed, and verified for essential properties (such as well-formedness and consistency) using the methods and tools of formal metamodeling. By design, Cy-PhyML is moving in the opposite direction to unified system design languages, such as SysML or AADL. Its goal is specificity as opposed to generality, and heavyweight standardization is replaced by layered language architecture and specification of explicit semantics.

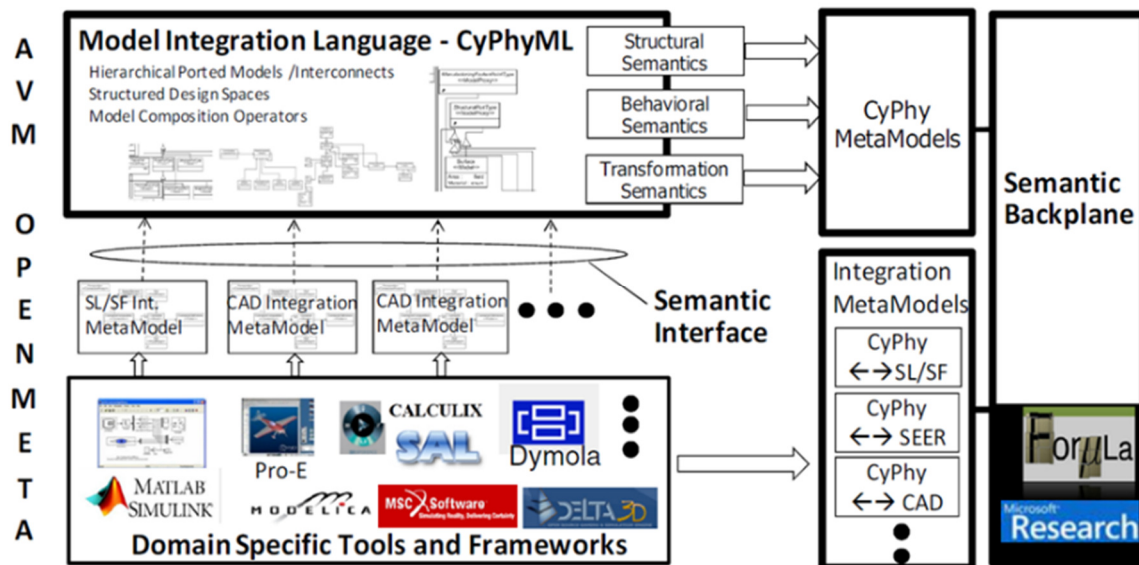


Figure 2-5 Method Integration Framework. - [45].

In Figure 2-5 it is observed as part of the Model Integration that a large suite of modelling languages and tools for multi-physics, multi-abstraction and multi-fidelity modelling are included; OpenModelica, Dymola, Bond Graphs, Simulink/Stateflow, STEP, ESMOL and many other software that are useful for analysis. In the end, CyPhyML model integration language provides the integration across this heterogeneous modeling space and the FORMULA - based Semantic Backplane provides the semantic integration for all OpenMETA composition tools [45].

Development: Tool Integration

Considering the approach proposed by [45] in which the Tool Integration Framework of the OpenMETA incorporate a network of model transformations that include models for individual tools and integrate model-based design flows, these model-transformations are used in the following roles:

- 1) Packaging. Models are translated into a different syntactic form without changing their semantics. For example, AVM Component Models and AVM

Design Models are translated into standard Design Data Packages (Figure 2-5, .ACM and .ADM files) for consumption by a variety of design analysis, manufacturability analysis and repository tools.

- 2) Composition. Model- and component-based technologies are based on composing different design artefacts (such as DAE-s for representing lumped parameter dynamics as Modelica equations, input models for verification tools, CAD models of component assemblies, design space models, and many others) from appropriate models of components and component architectures.
- 3) Virtual prototyping. Several test and verification methods (such as Probabilistic Certificate of Correctness - PCC) require test benches that embed a virtual prototype of the designed system executing a mission scenario in some environment (as defined in the required documents). We found distributed, multi-model simulation platforms the most scalable solution for these tests. We selected the High-Level Architecture (HLA) as the distributed simulation platform and integrated FMI Co-Simulation components with HLA.
- 4) Analysis flow. Parametric explorations of designs (PET), such as analysing effects of structural parameters (e.g. length of the vehicle) on vehicle performance or deriving PCC for performance properties frequently require complex analysis flows that include a number of intermediate stages. Automating design space explorations require that Python files controlling the execution of these flows on the Multidisciplinary Design Analysis and Optimization (OpenMDAO6) platform (that we currently use in OpenMETA) are auto-generated from the test bench and parametric exploration models (Figure 2-4).

The OpenMeta model and tool integration technology needs and infrastructure for creating and executing complex analysis flows. Based on “software-as-a-service” aspect of this development, it allows end users (individuals, research groups, and large companies) to repositories, analytic services and design tools to lower the costs, and exclude the high costs of acquiring and maintaining desktop engineering tools. In Figure 2-6 is presented the platform for executing the part of tool integration, according to [45].

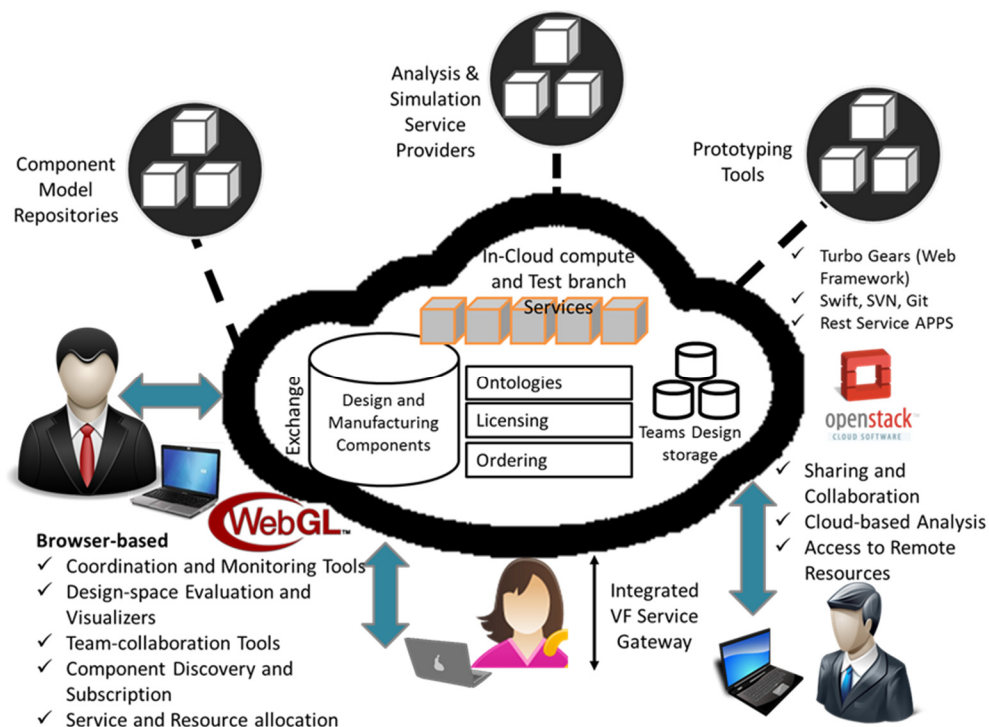


Figure 2-6 Tool Integration Framework. - [45].

With these fundamentals shown in Figure 2-6, it is clear that another matter needs to be addressed for this platform. The evolution of data is key in order to obtain better results and optimise the performance of this development. The importance of data management can result in ways to address customer needs and wants and improve designs in smarter ways. The next section includes the AI in the form of machine

learning approaches that helped to analyse the data and obtain useful knowledge for personalizing product designs.

2.3.2 Embedded Manufacturing Systems

Information and communication technologies form the bedrock upon which tomorrow's innovative solutions are built. Embedded systems and global networks are two major information and communications technology motors driving technological progress. Embedded systems already play a central role in today's lives, as are used to control many devices in common use today [46].

Embedded Systems are basically a computer system with a dedicated function within a larger mechanical or electrical system, constraints are often with real-time computing [47]. Those systems are the intelligent central control units at work in most modern technological products and devices. They typically operate as information-processing systems "embedded" within an "enclosing" product for a set range of device-specific applications. These "connect" with the outside world using sensors and actuators; allowing embedded systems to be increasingly interconnected with each other and the online world.

The difference between embedded systems and CPS, as discussed in [48] is that CPS describe an integration of computation with a physical process, then an embedded computer and network monitors and controls the physical process. CPS is about the intersection of the physical and cyber aspects of the manufacturing process or else, not the union. Thus, CPS means physical components and software (complete system), while embedded systems describe only the executable (computer) platform of the manufacturing process.

2.3.3 CPS and Data Analytics for Smart Manufacturing

In recent years, the use of sensors and networked machines has increased tremendously, resulting in high volumes of data known as big data being generated [49]. In that way, CPS, which exploits the interconnectivity of machines, can be developed to manage big data to reach the goal of resilient, intelligent, and self-adaptable machines. Boost efficiency in production lines for meeting customers' needs and wants is key in i4 principles, and since CPS are still under development according to [48], a proposed methodology and architecture described in [50] which consists of 2 main components: (1) the advanced connectivity that guarantees real-time data procurement from the physical world and information feedback from the digital space; and (2) intelligent data analytics, management, and computational capability that constructs the cyberspace. Figure 2-7 presents the value creation when combining CPS from an earlier data acquisition and analytics.

From the above framework, the smart connection plays an important role, hence acquiring reliable and accurate data from machines including components and customers' feedback telling the insides of the design that best approaches to their needs and wants. Here is where enterprise manufacturing systems intervene such as enterprise resource planning (ERP), manufacturing execution system (MES), and supply chain management (SCM). Data is obtained from those types of systems that update information in real time and provide a reliable inside of the product, from there all that collected data can be transformed into action [50].

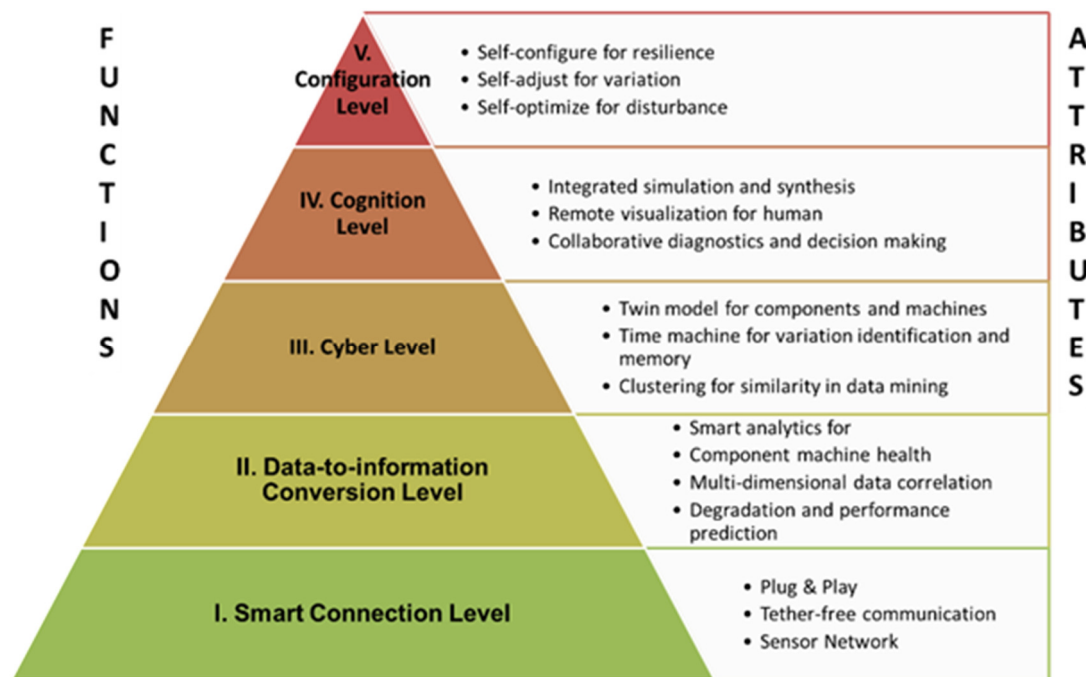


Figure 2-7 Architecture for implementing CPS [50]

i4 also describes the overlap of multiple technological developments that comprise products and processes. The purpose of this work is to provide a robust methodology to find possible solutions to fill the missing gaps that big data offers to individualistic manufacture (customized production). The next section gives a focusing on data managing (Cloud Computing) as well as big data environments is expanded as future directions for this literature review.

2.4 Big Data and Business Informatics for Industry 4.0

Considering business informatics according to i4 principles, companies need to tackle 2 factors: (i) the lack of an automated closed-loop feedback system that can intelligently inform business processes to respond to changes in real-time based on the inputs received (for example, data trends, user experience, etc.), and (ii) existing analytical tools cannot accurately capture and predict consumer patterns. These factors are due to business performance and the response to analysis outcomes, and

therefore it is essential to achieve real-time analytics to improve customer-business relationships as well as give customers an accurate product life-cycle in order to meet customers' desired usability of the product [51].

The use of digital models can be a possible way forward to address factor (i), such digital model needs to be capable of achieving automation in a closed loop. The vision of i4 is to utilize existing web-based technologies, internet marketplaces, and internet services where digital products are used as starting points to evolve better designs. Companies will have to be prepared for a digital transformation if they want to get the most out of i4 technologies. Cyber-security and data management will be critical problems to solve.

The use of intelligence for data businesses should also be in the collection of data, which can be viewed as an intelligent action. This is a possible solution to (ii). The intelligence in this way comes from an expert's knowledge that is integrated into the analysis, the knowledge-based methods used for analysis, and the new knowledge created and communicated by the analysis process.

Relevant to business informatics for i4 is prediction and analysis in customer needs and wants. Customer analytics for conducting business is concerned with analysing data, which also requires standard techniques such as data mining, statistics, intelligence data analysis, and machine learning.

Many statistical tools for achieving prediction in customer relations are not realistic for real-world applications [51]. In addition to this, [52] discussed real-time applications for IoT. The following tools and approaches were selected because of their promising results under smart environments and IoT according to [51] and [52].

- Data analysis using linear models, regularly performed in simple ways by utilizing linear functions, often does not represent the reality, as it is difficult

to describe real-world problems using such models. The problem relies on the use of linear statistics that involves numerous implicit assumptions about mutual independence between variables and normally distributed values. However, nonlinear models are more powerful for this.

- The hidden Markov models (HMM) [53] can be used for creating predictions on time-stamped events. Stochastic methods are represented by Markov models focused on the analysis of temporal sequences of separate (discrete) states.
- Using Bayesian networks [54] to analyse customer satisfaction is based on a graphical model, representing inputs as nodes with directed associations among them. Nonetheless, Bayesian networks do not provide all necessary levels of intuition, automation, or integration into corporate environments. Hence it is not very suitable for smart manufacturing. If Bayesian networks improve and find a way to be applicable to smart environments, it would enable accessibility to business users.

The term “big data” is a composite term, describing emerging technological capabilities in solving complex tasks. Highly acclaimed by industry analysts, business strategists, and marketing experts as a new frontier for innovation, competition, and productivity, big data leads to the new era of smart businesses, manufacturing, and virtualization of companies around the world.

Big data motivates researchers from fields as diverse as physics, computer science, genomics, and economics; it is seen as an opportunity to invent and investigate new methods and algorithms capable of detecting useful patterns or correlations present in big amounts of data. Analysing more data in shorter spaces of time can lead to better and faster decisions in areas spanning finance, health, and research.

A very simplified big data value chain includes at least 4 stages [55]:

- 1) Data is collected where it originates. During the data-generation stage, a stream of data is created from a variety of sources: sensors, human input, etc.
- 2) The raw data is combined with data from other sources, classified, and stored in some data repository.
- 3) Algorithms and analytics are applied by an intelligence engine to interpret and provide utility to the aggregated data.
- 4) The outputs of the intelligence engine are converted to tangible values, insights, or recommendations.

2.4.1 Role of Big Data Analytics in IoT

In the conception of i4 and smart manufacturing there was always this emerging topic known as IoT. What exactly does the IoT has to do with shifting the way businesses are made, products are manufactured, and services are given? The answer relies on the architecture behind the IoT that allows a wide range of controllers, sensors, appliances, and devices to be connected as part of the vast Internet [56].

Briefly what really motivates this work is how to bring together intelligent systems and automated decisions to execute them in many environments. The challenges that need to be considered are encountered in the following questions [52]:

- 1) How does an intelligent system effectively learn from data, and dissociate signal from noise?
- 2) How can an intelligent system integrate expert knowledge with observed patterns?

- 3) How can an intelligent system understand the context (Where, When, Who, Where) and act accordingly?
- 4) How can an intelligent system comprehend the consequences of and interference between different actions?
- 5) How does an intelligent system plan for causality that is not instantaneous, but takes place over time, across control iterations, and can fail?

The above questions can represent challenges not only for IoT, but as well for domains that similarly can take benefit of machine learning, AI, and expert systems. Nonetheless, the particular interest for exploration rises up for IoT domains because of the availability of big data inside of it. Furthermore, these aforementioned questions are linked to this work specifically in predicting customer needs and wants, since is relevant to know how an intelligent system learn from data and integrates this knowledge to put it into context applicable to individual needs when the design needs to be tuned accordingly. The next list gives a classification to decision systems, describing the required analytics, and different degrees of control capabilities, according to [52].

- **Reactive systems:** described as systems that take certain actions when a condition or criteria is met. The clearest example is the smart lighting systems that have sensors to detect the presence of persons in specific areas or rooms if the sensor does not detect a person, the lights remain off.
- **Knowledge-driven intelligent systems:** these systems try to capture the relationship between various domains, optimise decisions across these domains. Experts specify commonly the knowledge base, they might run a partial automated deep learning of cause and effect correlations within and across domains.

- **Visual Analytics:** these techniques facilitate the process of data analysis presenting relevant information displayed thoughtfully in a dashboard interface. Designed to boost peoples' decision process by presenting adequate information, and presenting it a cohesive and easily understandable manner.
- **Control and optimize:** these systems work in closed-loops in which decisions lead to instant actions, considering always the possibility that actions might fail meeting the optimisation goal. Attempting to optimise the behaviour of specific variables and consider cases as well when deciding action outcomes can fail, the control loop decision systems can also generate an action plan, execute it, observe the response in the closed-loop, and recover from a failure to meet a goal.
- **Behavioural and Probabilistic Systems:** in IoT human beings as users are an intrinsic part. Therefore, both become sources of data and means of control, using messages, suggestions, and incentives. Including human models as part of the entire IoT system, is what behavioural systems try to achieve and sometimes address stochastic (non- uniform) behaviour. Here is where fuzzy and probabilistic systems can incorporate non-deterministic models for decision-making, this stays on top and beyond failures.
- **Alerts and warnings:** end users provide decision logic in these systems. Then the information retrieved is used to interpret and classify the arriving data for raise alerts or make warnings. When users deal with large volumes of data, a kind-of automated predefined analysis to help detect or highlight situations of interest that might become critical.
- **Complex systems:** these systems are capable of understanding a context and interaction between many decision loops. The complex systems are also

capable of taking high-level decisions that span multiple dimensions within a single domain.

The above-discussed decision systems for IoT applications that have the most similar use to what we find applicable to our research line are the combination of visual analytics, control and optimise, and behavioural and probabilistic systems. Is central to our research interests to have a system capable of customising designs according to customer needs and wants, and here is where customers' behaviour needs to be fed into the system in order to recognize patterns, interact with customers, get the desired quality and design, and finally build the customised product tailor-made. The various aspects and applications of IoT enable or enhance the applicability inside the IoT by the integration of tools like visual analytics and optimisation for which machine learning, evolutionary algorithms, and AI, in general, have a natural way of solving these problems.

In the next section is presented the way AI can help giving solutions to the aforementioned IoT applications for smart manufacturing and enhance product design according to customer needs and wants.

2.4.2 Big Data Analytics Tools for the Smart Manufacturing Value Chain

As stated at the beginning of this section, for i4 and smart manufacturing processes dealing with large data storage, sharing data, processing, and analysing are becoming key challenges to computer science research. These challenges are (i) efficient data management and (ii) rapid time-critical processing requirements. Additional complexity arises from dealing with semi-structured or unstructured data. To understand massive amounts of data, advanced visualization and data exploration techniques are required [57].

The necessary features that need to be considered when big data are involved are categorized by two sources of data: (i) human-generated data, and (ii) machine-generated data. In specific, and for the purposes of this work, human-generated data includes needs and wants, amongst all types of data generated by people. Both present huge challenges for data processing. Big data cannot be defined by data volume alone. Complexity arises from the speed of data production and the need for short-time or real-time data storage and processing, from the heterogeneous data sources, from semi-structured or unstructured data items, and from dealing with incomplete or noisy data due to external factors. All these aspects render the analysis and interpretation of data a highly non-trivial task. It becomes even more challenging when data analysis and decision-making must be carried out in real time. Existing technologies for big data and machine learning [57] are presented in Table 2–1.

Table 2–1 Existing technologies for big data analysis and machine learning

Platform Name	Type of Analysis	Features
R Project	Statistical Analysis	<ul style="list-style-type: none"> • Combine statistical methodologies • Produce output to feed decision support systems • Process massive data from different sources
WEKA project	Data mining	<ul style="list-style-type: none"> • Flexibility for machine learning methods • Support the whole process of data mining, preparation of data, and statistical evaluation • Open-source software is written in Java • Support streamed data processing
KNIME	Data analytics	<ul style="list-style-type: none"> • Provides integration of new algorithms and tools and data manipulation • User interface allowing interactive exploration of analysis results or models • Continuously maintained and improved • Combined with powerful libraries such as WEKA data mining toolkit and R statistical language
Apache Mahout	Machine learning	<ul style="list-style-type: none"> • Provides machine learning algorithms that are scalable for large amounts of data

		<ul style="list-style-type: none"> • Many algorithms have been implemented for data clustering, classification, pattern mining, dimension reduction, among others
MATLAB	Data analytics	<ul style="list-style-type: none"> • Streaming algorithms perform stream processing • Machine learning toolbox • Hadoop enables MapReduce toolbox to work through the cloud • Cloud computing for extracting analysis and processing data without maintaining a data centre

In addition to the information presented in Table 2–1, there are some specific platforms that have applications for IoT according to [52] and [58]. Central to this work is to revise platforms and technologies that can cope with smart environments. These platforms need to have connectivity that can cope with IoT. The following list complements the existing technologies presented in Table 2–1.

- Apache Spark, developed by the University of Carolinas' Berkeley AMP Lab. This platform is an alternative to Hadoop and can perform in-memory computations. The Spark platform is a general engine for large-scale processing that supports Python, Scala, and Java. For certain tasks, it is up to 100 times faster than Hadoop MapReduce when the data can fit in the memory, and 10 times quicker when data resides on the disk. Recently, this platform was run on Amazon Elastic Map-Reduce [58].
- Microsoft Azure, Microsoft's response to big data analytics. This platform provides flexibility over the MapReduce by allowing users to have more control over the coding. Has a C#-like environment, and uses LINQ (a parallel language) and cluster execution. Part of its advantages is the debugging and development using Visual Studios as the tool for language interoperation [58].
- Google Cloud Platform & MillWheel, developed by Research Google. This platform is used for low-latency data-processing applications. The Google

Cloud platform includes many features for big data analytics and includes a machine learning application. Features of the machine learning approach include training and building models using the TensorFlow library [52].

The following list presents the requirements for an integrated platform for big data analysis based on the analysis discussed in [30], where several applications or case studies were considered. These requirements are considered as well for the existing technologies for big data and machine learning, which have been presented in Table 2–1.

- Functional requirements.
- Data integration: addressing problems from real-world application domains. Platforms must be capable of accessing multiple different data sources and dealing with inconsistent or noisy data.
- Statistical analysis: analysis of data can be simple or complex. Platforms must support different types of data analysis, including calculation of statistical key figures like quantiles or correlation coefficients.
- Interactive exploration: the platform has to support intuitive visualization for visual analytics and easy interaction with the data.
- Decision support: in addition to the analysis of data, the platform should also provide mechanisms for domain-specific data interpretations that are valuable for decision-making.
- Non-functional requirements.

- Scalability: the platform and its various constituents have to be able to handle huge amounts of data. All components must be designed in such a way that they can be deployed in a distributed computing environment.
- Near real-time processing: fast processing is the main requirement of some applications (use cases). The core platform must be able to support near real-time processing in combination with selected components.
- Resource efficiency: while keeping the objectives of throughput and speed, the system resources should be utilized efficiently.

All these requirements can also fall into three different research areas:

- Database Technologies: distributed databases, parallelism, and NoSQL approaches.
- Information Systems: design of an integrated platform with the scalability of all components and efficient usage of IT resources, making use of current system architectures (multi-core) and increased availability of main memory.
- Algorithmics: design of efficient algorithms for external memory, and algorithms fitting into the MapReduce paradigm or other parallelization patterns. Streaming algorithms are used for efficient processing of amounts of data that are so huge that scanning it more than once is infeasible, or for processing data that naturally arrives as an event stream.

Once requirements for big data are shown, it is necessary to see what needs are relevant to big data analytics. According to [32], the growth of the digital universe is expected to change from 898 exabytes to 6.6 zettabytes between 2012 and 2020, or

more than 25% a year, i.e., growth will double about every 3 years. In Figure 2-8 is presented the estimation made by the International Data Corporation.

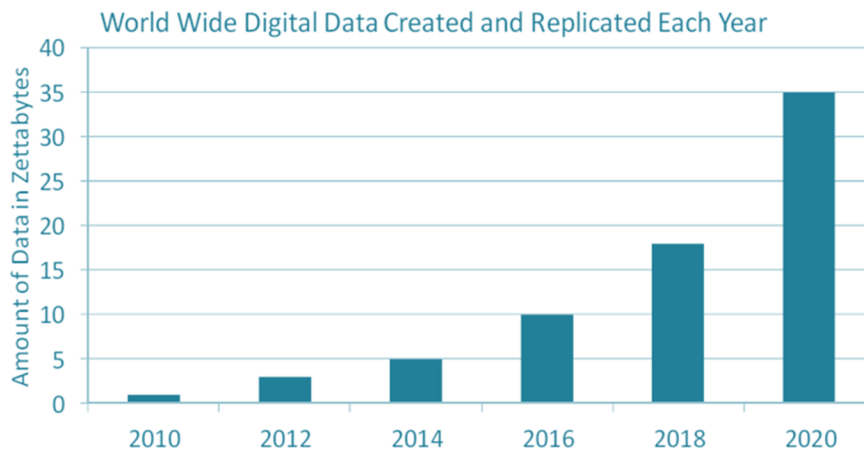


Figure 2-8 Expected growth in digital data from 2010 to 2020 [59].

Knowing that the growth of data would be constant nowadays, the question is, are the data analysed for useful information proper for instant usage? The value of a big data implementation will be judged based on one or more of these 3 criteria [33]:

- Able to provide more useful information;
- Able to improve the reliability of the information;
- Able to improve the timeliness of the response.

Thus, a big data application framework that meets the above criteria inevitably provides reliable, useful, and timely information, enabling a quick response by the data owner. If these criteria are not met, the big data is worthless.

Big data processing has become imminent for enterprises that wish to process a large amount of data mainly from social networks, the semantic web, sensor networks, geo-

based service information, patient information, and employee-based or transaction-based applications. These areas observe the quick growth of large data and need either timely analytics or batched processing. Thus, the challenge is to analyse and mine these big data to effectively exploit the information and improve efficiency and quality of service for consumers and producers alike. Computing capabilities of current multi-core microprocessors are unable to meet the data mining requirements to effectively mine the data on time, which means there is a need of parallel acceleration hardware, such as a graphics processor unit (GPU), to accelerate the data mining. MapReduce framework-based applications, such as Apache Hadoop and Drill (which are free and stable), are suitable for large-scale data processing.

2.5 Challenges Achieving Mass Customization

With the CPS-enabled i4 factory and big data advised design for agile manufacturing, the remaining challenge is how to achieve agile customisation, which is the focus of this section. For manufacturing, mass customization is the term used to describe the automated manufacturing of tailor-made products. Here, digital manufacturing and smart factory concepts are included. Where Direct Digital Manufacturing (DDM) happens all the way from the vertical integration, and the design of a product is passed on in a virtual form to the suppliers. The design can be passed on as a whole product or in several components. After that, each supplier contributes to adding specific data backwards to produce single components. Data then is passed to machines, and each part is produced directly from that data [60].

Production and many aspects of traditional manufacturing are affected by smart manufacturing. In the future, mass customization should start in the cloud. Since it can be difficult for a specific individual or team to handle all customizations, enabling public access for customers and suppliers to complete design and configuration along the value chain is a possible way forward. Fitting customers' needs and wants is located here, this takes place when customers place the order and deal with the

design and product configuration. ERP, such as customer relationship management (CRM), can be based on the cloud for historical orders, and also provide data analysis for self-aware properties. All the components involved in the integration framework for mass customization are shown in Figure 2-9 [61].

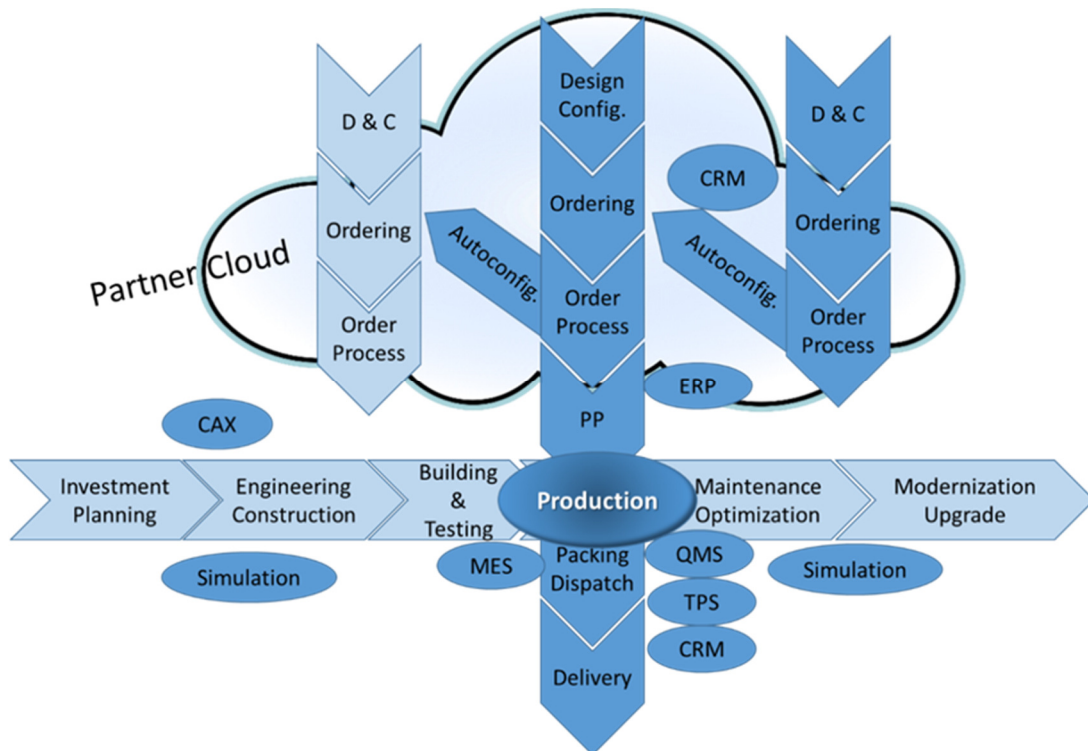


Figure 2-9 Horizontal and vertical integration in mass customization [60].

Components of the above figure encompass processing orders on the cloud, for which suppliers and customers collaborate at the moment of Design and Configuration (D & C), ordering, and process production (PP). Once the order is placed, it gets to processing for production stage, where just before producing, the simulation takes part in the following: planning investments, engineering construction, building and testing, optimizations, maintenance, and upgrades. Production integration continues until the package is dispatched and the order is delivered. The manufacturing software tools involved in this vertical and horizontal integration are as follows: CRM,

the Manufacturing Enterprise System, the Quality Manufacturing System (QMS), Computer-Aided Exchange (CAX), and the Transaction Processing System.

Stock, et al [17] discuss opportunities for sustainable i4 environments from a macro and micro perspective. These are summarized as follows:

- **Business models:** Smart data is a driver for new evolving business models in i4, enabling new services. In the long-term, sustainable business models are necessarily characterized by competitiveness. In the same way, selling the accessibility and functionality of products instead of selling only tangible products will be a leading concept [17].
- **Value-creation network:** In Figure 2-9, a value creation network as a crossed-linked cycle in i4 gives new opportunities for realizing closed-loop product life cycles and industrial cooperation. Having this cooperation of all parts on the cloud allows for efficient coordination of the material, product, and energy throughout the product's lifecycle as well as between different factories. The closed-loop life cycles also enable multiple use phases with remanufacturing or reuse in between [17].
- **Products:** Designing products under i4 principles is focused on realizing closed-loop life cycles for products by enabling the remanufacturing and reuse of specific products or by applying cradle-to-cradle principles. As the outcome of the manufacturing process, product quality can provide much insight on machine conditions via backwards reasoning algorithms [62].
- **Processes:** Data-driven and cloud-based technologies are key to achieving self-aware and self-learning machines. The design of proper manufacturing process chains by introducing data-driven techniques enable these characteristics. The importance of leveraging on additional flexibility and

capabilities offered by cloud computing is imminent, but adapting prognostics and health management algorithms to efficiently implement current data management technologies requires further research and development [62].

Considering what is addressed by [16] and [17] when trying to achieve mass customization, it is clear that the role of data, how is managed, and used represents a key challenge for further applications. CPS under the i4 vision will implement mass production and intercommunication through IoT, but mass customization needs to be designed in advance. However, it is often found that customers are not clear about their needs and wants [63]. Suddenly, how data is managed will lead to evolution of the innovation floor because the constant communication and linkage that IoT enables. In specific, data aims to move from manual settings to the automation and innovation of this process. Moving to an automated selection of design patterns and attributes is the purpose of this work, aimed to obtain the customized design of products.

For any company trying to address customer needs and wants, it is discussed in [64] that part of the main challenge is for the business to be responsive to the market speed, reason why is adopted the term “agile” since the company can improve its agility of product manufacturing by strengthening its ability of responding or controlling future market changes. Agile customization then is described as the specific task of adjusting and being responsive to customer needs and wants in such environment where individual desires change rapidly [65].

2.6 Summary

In this chapter, the main focus has been a review of the state of the art of smart manufacturing approaches and related research for the forthcoming Industry 4.0. Table 2–2 summarises the main ideas of reviewed research papers.

Table 2–2 Summary of the state of the art

Development	Authors/year	Main focus (ideas)
A High-Fidelity Temperature Distribution Forecasting System for Data Centers	Jinzhu Chen, Rui Tan, Yu Wang, Guoliang Xing, Xiaorui Wang, Xiaodong Wang, Bill Punch, Dirk Colbry (2012)	<ul style="list-style-type: none"> •Cyber-physical approach for temperature forecasting in data centres, which integrates Computational Fluid Dynamics (CFD) modelling, in situ wireless sensing, and real-time data-driven prediction. •Simulated temperature distribution and sensor measurements are then used to train real-time prediction algorithm. •CFD is a numerical tool that can simulate the future evolution of temperature distribution of data centres, often yields highly variable accuracy, poor scalability, and prohibitive computational complexity. •Use linear models as well to achieve real-time prediction. •Provide a well-founded methodology as well as the models used to train the GA •The approach can accurately predict the temperatures up to 10 minutes into the future, even in the presence of highly dynamic server workloads. •CFD models created for large-scale data centres typically have a coarse granularity and considerable error, this work has a better fit for minimum-scale data.
Cloud computing for industrial automation systems a comprehensive overview	Omid Givehchi, Henning Trsek, Juergen Jaspernite. (2013)	<ul style="list-style-type: none"> •Latest concepts of cloud computing technology for industrial automation focus •Growing of Industrial Revolution 4.0 based on intelligent production •Summary of all the authors, approaches and work done form the hand of cloud computing •Cases of cloud computing applied to automation, an architecture developed to improve information flow •Outlook the gap for cloud solutions in automation applied to lower levels •Control level achieve reliability and real-time issues

Collaborative systems for smart environments: Trends and challenges	Luis M. Camarinha-Matos, Hamideh Afsarmanesh (2014)	<ul style="list-style-type: none"> • Survey of trends and challenges for smart environments (modelling, design and development of collaborative systems). • Address paradigms like cyber-physical systems, Internet of Things, Internet of Events and Sensing Networks as supported technologies. • Related aspects: ambient intelligence, ambient assisted living, and sensing enterprise. • Areas of application: smart home, smart cities, intelligent infrastructures, intelligent transport systems and smart grid. • Highlights that modelling is a fundamental part of the development of future smart environments. • Point technical aspects like human-systems interaction, risks and security, technological basis, cloud computing, big data/data science.
Cyber-physical production systems: Roots, expectations and R&D challenges	László Monostori (2014)	<ul style="list-style-type: none"> • Cyber-Physical Production Systems relying on the newest and foreseeable further developments in computer science, information and communication technologies, and manufacturing science and technology lead to the 4th Industrial Revolution (Industry 4.0). • Industrial production of the future will be characterized by strong individualization of products under the conditions of highly flexible (large series) production, extensive integration of customers and business partners in business and value-added processes, and the linking of production and high-quality services leading to hybrid products. • Grid computing led to grid manufacturing, and similarly, cloud computing to cloud services to manufacturing. • Several acknowledgements' have driven to join Manufacturing and AI for learning and

		<p>prediction process now called Intelligent Manufacture Systems.</p> <ul style="list-style-type: none"> • Overview of the evolution Industry has led through the years and the future perspective of Cyber-physical systems. • Research and Development expectations towards CPS and CPPS are versatile and enormous: robustness, autonomy, self-organization, self-maintenance, self-repair, transparency, predictability, efficiency, interoperability, global tracking and tracing, etc.... • CPPS can be considered an important step in the development of manufacturing systems.
Global footprint design based on genetic algorithms - An "Industry 4.0" perspective	Guenther Schun, Till Potente, Rawina Varandina, Torben Schmitz (2014)	<ul style="list-style-type: none"> • Comparative study of network structures for optimizing costs in different scenarios. • Global footprint defined as the global distribution of production sites for a company. • Solution for unpredictable planning environment of manufacturing systems using GA's. • Simulation and virtualization reduce and optimize costs, improve decisions and solutions for a future production with accelerated development process. • Methodology for approaching the optimization and migration paths of production networks. • Optimization handles different scenarios, select the promising ones, GA's where helpful to obtain the best solution. • Improve the results of the production network design process and lead to further cost optimizations. • More analysis needs to be done to improve methodology.
Recent advances and trends in predictive	Jay Lee, Edzel Lapira, Behrad Bagheri, Hung-an Kao (2013)	<ul style="list-style-type: none"> • Acknowledge the concept of predictive manufacturing as manufacturing sector next transformation.

manufacturing systems in big data environment		<ul style="list-style-type: none"> • By embracing emerging technologies such as cyber-physical systems and advanced analytics manufacturers will improve efficiency and productivity. • Discuss the needs and technologies for predictive manufacturing systems in big data environment. • Give framework for predictive manufacturing and a cyber-physical model for enhanced predictive manufacturing system. • By implementing the prognostics and health management as well as analytical algorithms can accurately increase productivity. • Cyber-physical models integrated with simulation can continuously record and track machine conditions during several stages proposed.
Service Innovation and smart analytics for Industry 4.0 and big data environment	Jay Lee, Hung-An Kao, Shanhu Yang (2014)	<ul style="list-style-type: none"> • Manufacturing and new service trends on big data prediction to achieve high productivity through industrial virtualization and Industry 4.0 • Control machines to become self-aware and self-learning by managing together the whole interaction system • Assembly lines are highly automated and require new technology as well as intelligent systems to handle all the data • Cyber-Physical system is key between the physical world and cyber (computational) space, how the system interact with the machines to obtain optimal solutions • Prognostics and Health Management (PHM) Algorithm an is used and with clustering, it's set up the rules for how the system get the knowledge and adapt it to the changes through time
Smart Factory - A Step towards the Next	Dominik Lucke, Carmen Constantinescu, Engelbert	<ul style="list-style-type: none"> • Sketched Smart Factory approach

Generation of Manufacturing	Westkämper (2008)	<ul style="list-style-type: none"> •Decentralized information and communication to achieve real-time production •Highlight 3 challenges: Identification phase, Positioning phase, and Status knowledge. •The enabling of technology involves the concepts of embedded systems, (wireless) communication technology, automatic identification technologies, positioning technologies, federation platform, situation recognition, and sensor fusion. •Presented a functional architecture for a manufacturing enterprise which basically focuses on Product Data management, manufacturing execution, maintenance, education, and training functions. •Sensor technologies and integration of knowledge aim to increase the transformation of the factory. •Integration of heterogeneous information systems as horizontal and vertical reduces information deficits. •Based on Nexus Platform, vision the next generation real-time and context-aware production systems.
Survey of Recent Progress in Networked Control Systems	Ke-You You, Li-Hua Xie (2013)	<ul style="list-style-type: none"> •Provide a review of the state of art of Networked Control Systems, discussing various network conditions like minimum rate coding for stabilizing linear systems, network topology, and event-based sampling for energy and communication efficiency. •Properties of existing networks adopted in NCS since the development of the control technology in NCS, motivated by the type of networks used. •Evolution of control system technologies is reviewed. •Control technologies are affected by instrumentation for implementing control systems.

		<ul style="list-style-type: none"> • Discuss the minimum data rate problem for stabilization of linear systems over noiseless and noisy feedback channels, respectively. • Discuss a random down-sampling method to deal with the temporal correlations of the packet loss process. • Control of multi-agent systems which consist of multiple interacting linear systems is discussed. • Suggest research directions such as information transmission theory of NCS, performance control, network topology and data rate for multi-agent systems, cooperative control over uncertain large-scale networks and cyber security and fault tolerant control.
--	--	---

Summarizing those related works and developments leads to focus on the following aspects when facing i4:

- 1) The methodology that integrates collaborative systems, in this case, many researchers suggest that a well-funded methodology that integrates CPS, cloud computing, virtual designs and real-time analysis, is key to achieve a high productivity because the system at the end becomes self-aware and self-predictive among other properties that are suitable for study.
- 2) Decentralized intelligence, this paradigm comes along with the idea of keeping information and communication between the system components decentralized and by simulation and virtual design the manufacturing keeps improving, therefore optimization tool is used as well as control tools when setting the system scenarios.
- 3) Model-based integration, this approach requires significant future research effort. Many authors agree that modelling from the CPS is a big obstacle for

companies that handle big data, and profitable analysis obtained for prediction. Other focusing suggest tackling uncertainties within the data analysis. Tool integration and support from model-based systems and rapid construction of domain-specific toolchains is another suggestion from research.

- 4) Experimental research to validate scientific results of the theoretical work is also what authors suggest. Validation and implementation of these approaches, because with the fast rhythm of acquiring knowledge and developments, what is trending now, in few more years will not be the same. When launching these projects like smart manufacturing and Industry 4.0, companies need to stay one step beyond and put effort into innovative resources, in order to get better results.

Collaborations to trigger necessary technology for Industry 4.0 are found in many study cases included in this section, many of those applications show how CPS can interact with the manufacturing process or system. Having a variety of high-tech machinery is not all, integration methodologies to minimize error, and to make interactive the system with human support is also key. By keeping a simple and useful interaction of virtual-physical-human part of each process is essential. Visualization tools can represent for human tasks a high-value development, in order to constantly supervise the process and minimize performance errors.

Design attributes can play the role of customer needs and wants and as discussed in chapter 1, there is a way of informing the manufacturing process what attributes are more desirable for individual users. It is seen that with the CPS-enabled i4 factory and big data advised design for agile manicuring, the major research challenge remaining is how to achieve agile customisation, as described earlier in section 2.5 in this chapter. This is therefore the focus of this thesis. The next chapter will hence analyse and develop suitable methodologies for achieving agile customisation for i4.

Chapter 3 Methods for Attribute Prediction Using Smart Design Principles

In this chapter, the methodologies that according to the literature review may lead to a solution to the missing gaps in the i4 concept are presented. In this sense, the solution means finding an effective way of individualizing product production and still being able to continue with the benefits that mass production offers. Now, according to the reviewed literature, the possible directions are considering a full integration of technologies (same as concepts) that can cope with all the necessary tasks a smart factory require. Many of the challenges for i4 and smart factories is the integration of information and communications technologies, CPS, and IoT. The integration of these technologies should be suitable for automation and a predictive closed loop. In this chapter, Smart Design Principles are considered to be equal to i4 principles that were already discussed in Chapter 2, but extended from the i4 view since in this thesis the smart design is a key aspect when addressing customer needs and wants for i4 environments.

Another perspective suggests considering the potential as well that data brings to manufacturing and design in the digital age. In this sense, data analytics bring a complete focus to find a possible solution for customization. The use of AI and machine learning approaches are key in this process. As discussed in [66], there is no single algorithm or approach that works better than the others on a general basis. Therefore, for each problem, an appropriate algorithm needs to be assigned. The selected algorithm needs to provide the desired performance and results for each specific application.

The following sections show the methods that unite these two perspectives considering the perspective that designs can be customized digitally and improved

according to individual needs. The first section presents the CPS integration and gives three different developments that help us to set up a starting point for proposing a framework. The second part includes the machine learning approaches as part of the data analysis and meet customer needs and wants for obtaining better and more informed designs. The third part included in this chapter discusses the smart design and automation approach since this is one of the key points when addressing customer satisfaction. Finally, a summary is presented with the motivation of choosing these approaches.

3.1 Hypotheses to Set the Scene

Based on the reviews of the state of the art and the problems that need to be solved, this section presents research hypotheses to set the scene. The first hypothesis addresses the general approach to develop a framework to automatically predict the design attributes that best reflect what customers need and want in a product, looking for research and technical evidence to support that such framework can be developed.

Hypothesis I

It is possible to develop a framework capable of automatically predict the design attributes that best reflect what customers need and want in a product.

Hypothesis I is expected to be clarified with the following research questions:

1. How can a generalized framework that automatically predicts the design attributes that best reflect what customers need and want in a product be developed?

2. Which approaches can effectively help to predict the design attributes that match customers' requirements?
3. How can products be designed efficiently?

Hypothesis II

It is possible to obtain a model capable of accurately predict customer needs and wants for at least 85% of classified design attributes.

Hypothesis II is expected to be clarified with the following research questions:

1. How can design attributes be used to provide meaningful insight of customer needs and wants?
2. Which artificial intelligence approaches can be tested to obtain classification models that best represent customers' needs and wants?
3. Can a classification model that scores less or close to 85% be reliable and used for prediction?

Hypothesis III

It is possible to identify effective ways of achieving customization for i4 and smart manufacturing.

Hypothesis III is expected to be clarified with the following research questions:

1. What are the identified challenges to be tackled for making effective the customization under i4 and smart manufacturing principles?

2. Which possible ways for achieving customization are effectively used in any current stage for i4 and smart manufacturing?
3. Do i4 and smart manufacturing deal with customization in particular ways different to other approaches already existing, tested, and working?

The following sections will present methodological considerations that may be used to address these hypotheses. Then a focussed framework and implementation approaches will be developed.

3.2 Artificial Intelligence for the Smart Manufacturing Value Chain

The current stage of manufacturing needs effective solutions to overcome the challenges that the extensive use of internet and information brings to global markets, and also the requirements of today's' customer are not the same as previous stages. Even though AI methods and same algorithms had been exploited and used for decades, the constant development of applications and implementations of AI in day-to-day life had left a solid foundation on how to actually benefit from it. The difference in today's applications is that implementations of AI are performed in more powerful computers, and algorithms had been trained in larger datasets [67]. In terms of functionality, algorithms are becoming "smart" because of the cognition aspect and the fact that help humans to make better decisions; the cognition aspect involves the process of acquiring knowledge, an example is found in machine learning where an algorithm is trained to recognize new patterns using the deductive technique. Now, if AI and machines had brought a better understanding of how to manufacture products and give services, the potential for addressing customer needs and wants for product design are huge.

The value of decision-making in i4 and smart technologies is central when it comes to effectiveness. IoT in i4 perspective is all about decision informatics and adopt the recent advancements of technology like processing (real-time analytics), sensing (Big Data), learning (deep learning/machine learning), and reacting/adapting (real-time decision-making) [68]. The key drivers for i4 are IoT, real-time decision-making (RTDM) and AI [59]. These technologies enable prediction (speech and synthesis), recognition (voice and video), and understanding behaviour (social-media) technologies to improve ineffective applications for i4 [69].

What is inside RTDM is a system of integrated computers that need to perform critical decision-oriented functions as part of the so-called decision informatics [70]. In the framework illustrated in Figure 3-1, there are four technologies (sensing, processing, learning, and reacting) for real-time decision informatics [69]. Smart designs will be always sensing, processing, reacting, and learning as part of a closed-loop.

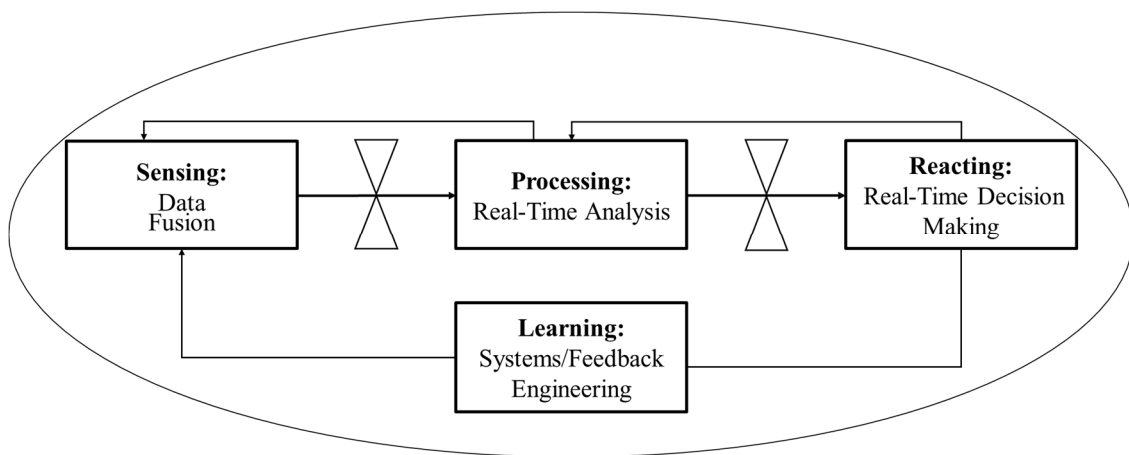


Figure 3-1 Framework for real-time decision informatics [69]

The idea for Figure 3-1 is that from a systems engineering perspective, RTDM determines:

- Data to be sensed, collected and fused from possible various data sources.

- Data analysis or processing for obtaining needed information.
- Reactions required to make informed decisions, decision-making and communications included.
- Learned knowledge to support future decisions and understanding of them.

This feedback loop from Figure 3-1 helps as well to refine all the steps involved (processing, sensing, learning, and reacting), but also includes visualization and management of data, mining, wisdom (reliability, quality, pattern analysis, fuzzy logic, AI, etc.), and knowledge [69]. How effective this framework is for specific problems lies on how relevant the models are for describing the problematic, since AI and learning steps are not just about speed but involve the analytical part as well. The best way for integrating products, operations, and processes in system engineering is a holistic approach in order to be able to adapt to changes since the aim for i4 is to have human-centred systems and intelligence-oriented as well.

However, there are certain conditions and technologies that enable digital manufacturing; these are novel materials, cloud computing, and smart robotics. How efficient and effective these technologies are based on their general objectives is presented in Table 3—1 [69].

Table 3—1 Digital manufacturing enabler technologies

Enabler	Methods	Objectives	
		Efficient	Effective
Cloud computing	• Software: unlimited, simulation, algorithmic	✓	✓
	• Hardware: unlimited, scalable	✓	✓
	• Cost: pay-as-you-use, cybersecurity concerns	✓	✓

Novel materials	<ul style="list-style-type: none"> • Creation: big data analytics, decision informatics • Technologies: graphene plasmonics, smart sensing • Cost: toxicity, environmental impact 	✓ ✓ ✓	✓ ✓ ✓
Smart robotics	<ul style="list-style-type: none"> • Software: digital designs, smart controls • Hardware: smart robots, 3D printing • Evolution: cheaper, more efficient, more distributed 	✓ ✓ ✓	✓ ✓ ✓

Digital designs are considered a key part of smart robotics (CPS) but as well the counterpart cloud computing also highlights the use of AI and simulation. These concepts embedded inside other technologies and approaches are taking part in the big picture that i4 represent. In the end, smart technologies are driving the 4th industrial revolution thanks to the state of maturity that many technologies had reached. The objectives of digital manufacturing are to make products efficiently and effectively, thanks to cloud computing, novel materials, and smart robotics and each methodology are now becoming a reality.

In AI the aspect that is getting most of the attention of researchers according to [67] is machine learning and is largely attributed to the fact that is simple to use, provides more insight from big data sets, and gives computers the ability to learn without being specifically programmed. Machine learning has evolved from the basics of pattern recognition to the construction of algorithms able to learn from data and make predictions. Algorithms can build models from sample data inputs, and this is useful when dealing with custom-design problems because it gives the opportunity to produce reliable, uncover hidden insights, and repeatability on decisions and results allowing the possibility for automation plausible.

The known approaches for AI learning include [69, 71]:

- Learning decision trees: the predictive model is obtained from a set of decision rules known as trees.
- Association rule learner: used for discovering useful relations between variables in large datasets.
- Artificial neural networks (ANN) learner: inspired by biological neural networks functionality.
- Deep learning: multiple hidden layers in ANN.
- Inductive logic programming (ILP): makes a uniform representation for background knowledge, input examples, and hypothesis.
- Support Vector Machine (SVM): used methods for supervised learning for classification and regression models.
- Clustering: observations assigned to a set or subset called clusters
- Bayesian networks: graphical model or belief network that represents a set of aleatory variables and the conditional independences between them.
- Reinforcement learning: based on the possible way an agent might take actions or decide based on objective function for a long-term target.
- Representation learning: aim to discover better representation of inputs.
- Similarity and metric learning: identify similarities or distance metric functions to predict if new inputs are similar.

- Genetic Algorithm (GA): a heuristic method for searching that mimics the process of natural selection.
- Rule-based machine learning: in this method is identified, evolved, or learnt rules to manipulate, store, or apply knowledge.

All the above-mentioned approaches for machine learning find an active field for applications inside manufacturing environments. AI and learning methods combined with cloud computing, as seen in Table 3—1, gives the opportunity to access unlimited source of processing and storage power, which additionally can be reconfigured and these features are becoming effective, faster, and cheaper [59]. The ultimate goal for machine learning and AI is to find the optimal solution, and when an application reaches desired results, successful decision-making can be obtained; this makes a positive impact on how effective a business can be. Under i4 and smart manufacturing perspective, AI works better when paired with humans, since these methods will only help people to understand the faced-problem, and to make the best decision based on the inputs and set of rules given to the algorithm. At the end of the day, AI becomes a key enabler for i4 allowing to reach the vision of Smart Factory - self-awareness, self-learning, self-control, self-adaptive, and self-organized processes.

As discussed previously, AI will not take a full part in building product-design and product manufacturing. It is best to have a methodology or concept inside AI that can help customers and designers to make the best decision, but considering users' needs and wants. For this, we present in the next subsection what are the predictive approaches to address customization.

3.2.1 Predictive Models to Address Customer Needs

In general, the idea of personalized products/services is to tailor features to known needs and wants of individual users. With this known features the information can be

used and/or stored in a user model to extrapolate which items like services, products, or units of information should be shown to another user, and make predictions based on known behaviour [72]. These approaches are known as recommendation technologies, which mainly use AI to support the identification of items to recommend or show to each user [73]. This recommendation technologies allow users as well to identify products and services that correspond best to their needs and wants. A basic architecture of recommendation system is depicted in Figure 3-2 [72].

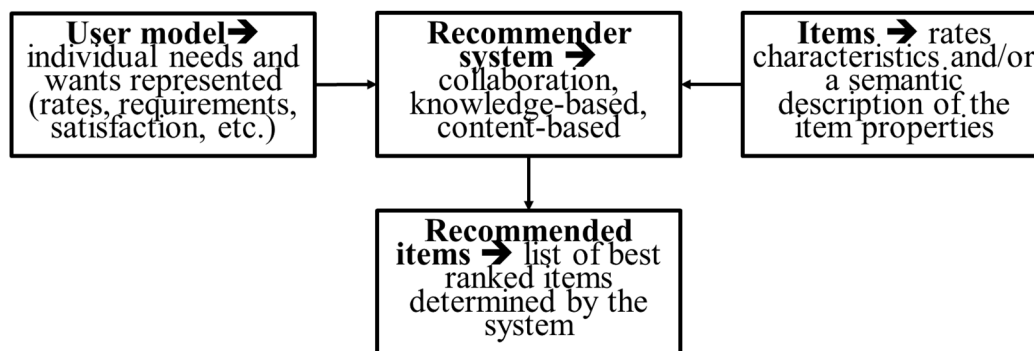


Figure 3-2 Basic architecture of a recommendation system [72].

In Figure 3-2 is observed how the user-related preferences are saved in the user model. Rates and additional semantic information are characterized in Items section. What is presented to the user in form of recommendation is the item catalogue to derive a ranked list of options. Here the recommendation system mines or exploits the information of the user model.

Recommendation system techniques have been widely used for online businesses, in specific the collaborative filtering approach [74] that represents one of the most used recommendation technique. There are many applications of this technique find in [75-78], and roughly consist of calculating common product features and recommend it to new customers during the design process. Nonetheless, the limitations of this approach lie on a term described in [79] as a cold-start problem, this means that no meaningful recommendation will be suggested to a user because of the lack of initial

ratings [80]. The cold-start problem is encountered in two situations: 1) when a first-time user interacts with the system and no ratings are registered to provide individual predictions, and 2) when new items are added to the system and had not been rated, therefore cannot appear in the recommendation list [81].

Although, techniques like content-based filtering and content-based recommenders are commonly used to address individual needs and wants; for this work, we decided to focus on the aspect of knowledge-based approaches, i.e. an expert system approaches that perform automated reasoning and knowledge (documents, media, inputs, etc.) to be leveraged by humans [82]. i4 is all about knowledge-driven technologies and it is more suitable for smart environments to have information pre-processed for better prediction. In knowledge-based techniques, information about users and products/services are used to perform reasoning in order to make recommendations on how items meet customers' needs and wants [83]. In knowledge-based techniques, recommendations are not based on user ratings, therefore the cold-start problem is not encountered [81].

Similar to knowledge-based recommendations, it is found in [84] an AI algorithm to mine data and generate knowledge on user needs and wants to match specific products by using machine learning approaches, in this case, Classification Based on Association (CBA) approach. The outputs obtained in the algorithm (knowledge), are then used to generate recommendations to new users. Another approach for knowledge acquisition is presented in [85] but this application is used when dealing with large and complex databases, and the author gives a solution to the bottleneck problem of data acquisition. One last interesting case for knowledge-based techniques is found in [86], in which the social aspect of customizing products is combined with social media by connecting the shopping experience through a common platform, where part of the inputs for customising the product are taken from social media

interaction. Once a person places an order, friends can get notified and the purchase is recommended to them, reaching more recommendations per item.

All these aforementioned approaches and techniques prove the fact that customization has evolved from simple applications for online business to the current state of customization and set the foundations for smart technologies and i4 environments. The tendency for personalized-product-designs is to allow users to get involved in the process of building their own products and share knowledge/experience with manufacturers in this process. This exchange part is pretended to be covered by IoT and Cloud Computing in i4. Recommender techniques for sure are helping to alleviate the state of confusion for customers to not know what they need and want, so the process of selecting an individual design can be facilitated by the use of AI. Using these algorithms and knowledge-based techniques inside design for manufacture results in a new paradigm shift called smart design.

Exploring all the possible ways new technologies and concepts can make i4 successful, comes the part of including human emotions to train algorithms to actually make decisions based on individual needs and wants. The next subsection discusses the affective design approach.

3.2.2 Affective Design for Mass Customization

Amongst many strategies that incorporate human emotions for addressing customer needs and wants and increasing the competitiveness of products, we have the affective design approach [87]. This concept implicates from one side a customer-oriented product design, and from the other, a manufacturing process that takes full account of customer needs and wants integrating several affective factors of individual users into the product-design process [88]. The combination of sentiments, emotions, and attitudes towards a specific product can be turned into design parameters used to meet the requirements of individuals as the custom design [89].

This useful approach involves AI, design, and psychological approaches for mining keywords that describe best the sentiment of each individual towards a product, then use this information as inputs to improve the decision-making of customizing a design.

Affective design comes from a technique proposed by [90] called Kansei engineering, that translates customer subjective impressions about a product, into design elements that can then be used to tune individual design for meeting customer needs and wants.

There have been many successful applications using affective design/engineering for customizing products. For this work, it was decided to focus on the ones that utilize learning and AI approaches to fulfil customer desires. Study cases found in [4, 89] discusses the use of AI to code affective needs and wants using the Kansei technique for obtaining useful attributes to associate with design parameters. Other developments presented in [81, 91] suggest the inclusion of virtual platforms in their methodologies, this helps customers to select the best product design for vehicles and the virtual platform is used as guidance through all the customization process.

Based on previous research regarding affective design, the idea of subtracting meaningful knowledge for training machine learning models and being able to make predictions, suggests that is important to create a set of rules based on specific products. This means that for every product design process it needs to be specific rules associated with personal requirements, but in [81] is discussed the possibility to pre-process raw data into training datasets, for which decision trees find a natural application since the model can be evaluated and refined iteratively until the desired confirmation is reached. This approach is in line with our work and most of the analysis obtained targets customer needs and wants in a predictive way, aiming to give a complete analysis when recommending personalized designs under i4 principles.

Customization using affective design approach can utilize many learning techniques for making decisions, but since classification techniques can easily associate known

attributes or design elements to new data for predicting what an individual might select, we decide to focus on this approach. In this context classification helps for decision-making because a known structure can be generalized, then as mentioned before, apply this to new inputs [92]. An example found in [81] of classification framework considering affective needs for making predictions is shown in Figure 3-3.

In Figure 3-3, the term design element de_i describes any customizable product part, and each de_i is characterized by a set of design parameters dp_{ij} or attributes that represent each element to the particular impression (shape, color, texture, etc.). Therefore, de_i is represented as a set of design parameters $de_i = [dp_{i1}, dp_{i2}, \dots, dp_{in}]$, and each parameter dp_{ij} has a set of possible values. It is presented the example of a round product made of aluminum, the material selection is represented by $dp_{11} = \text{material}$ for any given de_1 , in which a set of possible options can be selected (alloy, copper, steel, etc.). Based on this design elements, users comment their opinion of each presented element, this data is collected to build a classification model for each design parameter. Here classification is used to identify hidden relationships between each design parameter and customer affective needs. If the classification model represents an accurate value, predictions can be made of the specific design parameter that satisfies the affective need of new customers.

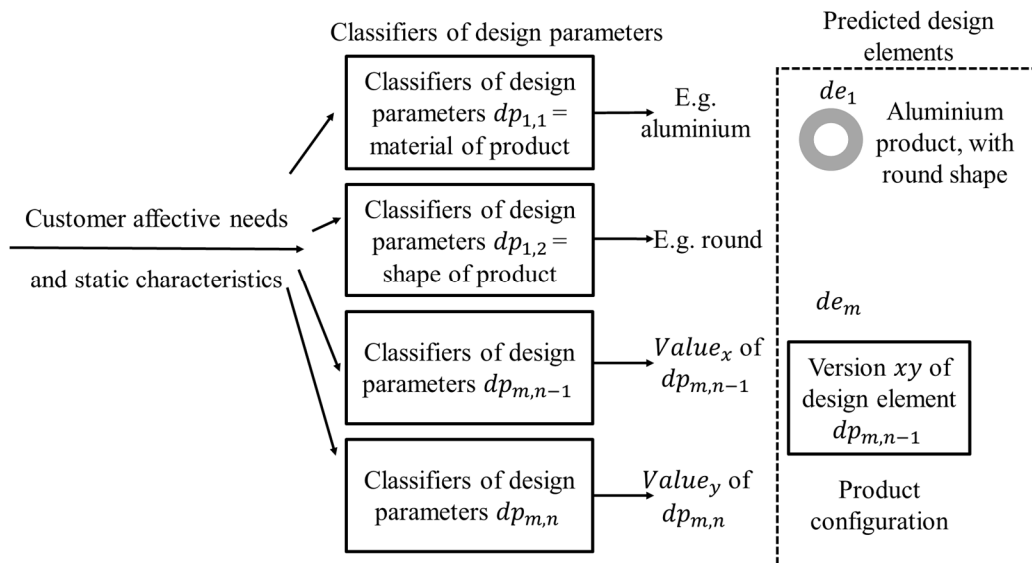


Figure 3-3 Prediction process of product configuration for new customers [81].

Ideally, many techniques can be adapted and tested to achieve customization. The interest and main objective for this work is to put together a methodology that encompasses the i4 principles in order to close the gap between current product design processes, to customized product designs manufactured in a smart way.

3.3 Machine Learning Based Approaches

To tackle affective design for mass customization, machine learning as a means of an AI approach may provide a powerful tool. Smart design concept under i4 principles aims to deal with large quantities of data in digital environments, and many studies focus on intelligent tools that help digital designs cope with RTDM and customer satisfaction. Considering this, AI and novel machine learning techniques, suit perfect for extracting hidden patterns from data [93], and also have a huge potential to provide a clear improvement of many transformation processes, as well as improve services by providing reliable insight into what customers really need and want.

In this work, it was not only necessary to predict customer needs and wants but to focus also on how to improve product-design according to individual desires. For this, we decided to include a recommender approach that was able to use the outcomes collected from the classification model, and then use it to recommend the best selection from the design elements suitable for prediction. However, it is first necessary to understand the characteristics of the data to find the most suitable method according to data inputs [71]. A good understanding of the dataset is crucial to this choice and the eventual outcome of the analysis. Many of the algorithms developed so far are iterative, designed to learn continually and seek optimized outcomes.

We also considered a business informatics perspective that encompassed i4 principles. In this perspective companies need to tackle 2 factors:

- (i) the absence of an automated feedback closed-loop method that can inform business processes in a smart way how to respond to changes in real-time based on the inputs received (data trends, user experience, etc.) and
- (ii) existing analytical tools cannot accurately capture and predict consumer patterns.

These factors are due to business performance and the response to analysis outcomes, and thus it is essential to achieve real-time analytics to improve customer-business relationships as well as give customers an accurate product life-cycle in order to meet customers' desired usability of the product [51].

We previously discussed in [27], that the use of digital models can be a possible way forward to address factor (i) since such digital models need to be capable of achieving automation in a closed loop. The vision of i4 is to utilize existing web-based technologies, internet marketplaces, and internet services where digital products are

used as starting points to evolve better designs, i.e. the existing digital products are based on designs previously utilized as manufactured products, where the idea is to use these product designs to evolve better ones by addressing customer needs and wants.

The use of intelligence for businesses focused on data (data businesses) should also be in the collection of data, which can represent an intelligent action. This is a possible solution to (ii) [27]. The intelligence in this way comes from an expert's knowledge that is integrated into the analysis, the knowledge-based methods used for analysis, and the new knowledge created and communicated by the analysis process.

The next subsections show the approaches used to analyse the data for predicting, pattern-detecting, and selecting suitable design attributes according to customer needs and wants. The subsections are organized as follows, cluster analysis used for detecting pattern and customer behaviour, classification analysis used for building predictive models and detecting significant attributes, and feature selection for determining significant design attributes and narrow down options for recommending design features.

3.3.1 Clustering Analysis

The use of cluster analysis in this work is attributed to the research found in [49], where predictive manufacturing methods are introduced for smart environments. In this research is found the idea of transforming processes assets' information to predict the health condition of individual machines, give maintenance, and take actions when needed, by using machine learning and specifically cluster tools to analyse the data. Cluster analysis was used because of the visualization tools, and health information (i.e., current condition, remaining useful life, failure mode, etc.) were successfully displayed in charts like fault map, radar, health degradation curves, or risk charts.

Here it was discussed the idea of including self-organizing maps (SOM) for performing the analysis.

Cluster analysis generally refers to a wide spectrum of methods that try to subdivide a dataset X into c subsets (clusters), which are partitioned in pairs, all nonempty, and X is reproduced via union [94]. After this, the clusters are designated a hard c –partition of X . The observations inside the data will have a membership in every cluster, the memberships close to unity can be interpreted as high degree of similarity between the sample and a cluster, and at the same time memberships close to zero denote minimum similarity between the sample and that cluster.

The intention using this approach for predictive analysis of customer needs and wants is to determine significant patterns, features, and properties inside the data that should be considered for specific individuals that match specific categories identified on the X clusters. For this, each observation inside the data can present several hundred dimensions, the variety of structures is without a bound. Here is clear that (i) no clustering measure or criteria of similarity will be universally applicable, and (ii) selecting a specific criteria is at least partially subjective, and therefore open to question [95].

In order to explore the capabilities that cluster analysis bring to this work, in the next subsections are included the main approaches used for predicting customer needs and wants.

Self-organizing Maps

To realize the i4 principles, full integration of CPS and powerful tools for optimization, clustering, modelling, selection, and prediction, is crucial for a complete analysis [49], [7]. The use of adaptive learning and data mining algorithms creates a knowledge base representing the scenario performance when either the characteristics (qualities

or features: colour, type, weight, etc.) of a product needs to be considered, or its attributes (characteristics to be associated to: specific brand, group of objects, etc.) need to be personalized. Then, those mechanisms can be automatically populated. The knowledge base will eventually be able to grow with new data to enhance its capability of representing complex working conditions that happen in real-world scenarios.

A SOM is a type of ANN that is trained through unsupervised learning, i.e., clustering. A SOM is made up of neurons (nodes), each with an associated weight vector. It is used in dimensionality reduction problems. Through adjusting the neurons and the associated weight vectors, it can produce low-dimensional cluster representations (2D map) of a set of high-dimensional input data.

The obtained map is a $N \times N$ space, where the data are scattered and arranged. The number of neurons is set as the square of the map. The function can be summarized in 4 steps [96]:

- 1) Initialization: all connection weights of each cluster are initialized.
- 2) Competition: for each input pattern, the neurons compete against each other to win this input. The neuron that adapts its value closest to the input wins. The discriminant function can be defined to be the squared Euclidean distance between the input vector x and the weight vector w_j for each neuron j , as follows:

$$d_j(X) = \sum_{i=1}^D (X_i - W_{ji})^2. \quad (1)$$

- 3) Cooperation: once a winning neuron has been selected, this neuron then creates a neighbourhood located close to the previous winner. Therefore, the

winning neuron creates a neighbourhood with other neurons to cooperate and win future inputs. If S_{ij} is the lateral distance between neurons i and j on the grid of neurons, we define a topological neighbourhood $T_{j,I(x)}$, where $I(x)$ is the index of the winning neuron and σ is the size of the neighbourhood, which needs to decrease with time:

$$T_{j,I(x)} = \exp \left(-\frac{S_{j,I(x)}^2}{2\sigma^2} \right). \quad (2)$$

- 4) Adaption: this last stage is when each neuron creates a neighbourhood or becomes a member of a neighbourhood and self-organizes so that the feature map between inputs is formed. The equation that describes the appropriate weight update is as follows:

$$\Delta W_{ji} = n(t) \cdot T_{j,I(x)}(t) \cdot (X_i - W_{ji}). \quad (3)$$

For every step, all neurons adapt their weights to the current input, but not as much as the winner neuron and its neighbourhood. Visualization of the map presents, in this way, every neighbourhood. Each neighbourhood value will be suitable for approximation values that have been ordered and shaped.

Fuzzy clustering

Cluster approaches can be applied to datasets that are qualitative (categorical), quantitative (numerical), or a mixture of both. Usually, the data (inputs) are observations of some physical process. Each observation consists of n measured variables (features), grouped into an n – dimensional column vector $z_k = [z_{1k}, \dots, z_{nk}]^T, z_k \in R^n$ [97].

The N observations set is denoted by $Z = \{z_k | k = 1, 2, \dots, N\}$, and is represented as a $n \times N$ matrix:

$$Z = \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1N} \\ z_{21} & z_{22} & \cdots & z_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ z_{n1} & z_{n2} & \cdots & z_{nN} \end{pmatrix}. \quad (4)$$

Clustering techniques can be categorized depending on whether the subsets of the resulting classification are fuzzy or crisp (hard). Hard clustering methods are based on classical set theory and require that an object either does or does not belong to a cluster. Hard clustering means that the data is partitioned into a specified number of mutually exclusive subsets. Fuzzy clustering methods, however, allow objects to belong to several clusters simultaneously, with different degrees of membership [97]. Fuzzy clustering assigns membership degrees between 0 and 1 that indicates their partial membership. Cluster partitions are vital for both cluster analysis and identification techniques that are based on fuzzy clustering.

Most analytical fuzzy clustering algorithms are based on the optimization of the basic c-means objective function or some modification of the objective function. The optimization of the c-means function represents a nonlinear minimization problem, which can be solved by using a variety of methods, including iterative minimization [98]. The most popular method is the simple Picard iteration through the first-order conditions for stationary points, known as the FCM algorithm. Bezdek [95] proved the convergence of the FCM algorithm. An optimal c partition is produced iteratively by minimizing the weighted within the group sum of the squared error objective function:

$$J = \sum_{i=1}^n \sum_{j=1}^c (u_{ij})^m d^2(y_i, c_j), \quad (5)$$

where y_i is the dataset in a d -dimensional vector space, n is the number of data items, and c is the number of clusters, which is defined by the user. Furthermore, $2 \leq c \leq$

n, u_{ij} is the degree of membership of y_i in the j th cluster, m is a weighted exponent on each fuzzy membership, c_j is the center of the cluster j , and $d^2(y_i, c_j)$ is a square distance measure between object y_i and cluster c_j .

The following steps were used inside MATLAB for the fuzzy c-means algorithm.

- 1) Input: c = centroid matrix, m = weighted exponent of fuzzy membership, ϵ = threshold value used as the stopping criterion, $Y = [y_1, y_2, \dots, y_n]$.

Output: c = update centroid matrix.

- 2) Randomly start the fuzzy partition matrix $U = [u_{ij}^k]$
- 3) Repeat
- 4) Calculate the cluster centres with U^k :

$$c_j = \sum_{i=1}^n (u_{ij}^k)^m y_i / \sum_{i=1}^n (u_{ij}^k)^m. \quad (6)$$

Update the membership matrix U^{k+1} using

$$u_{ij}^{k+1} = 1 / \sum_{k=1}^c \left(\frac{d_{ij}}{d_{kj}} \right)^{\frac{2}{m-1}}, \quad (7)$$

where

$$d_{ij} = \|y_i - c_j\|^2, \quad (8)$$

until $\max_{ij} \|u_{ij}^k - u_{ij}^{k+1}\| < \epsilon$.

5) Return c

In the next section, the classification approaches are presented in order to give a more complete analysis for predicting customer needs and wants.

3.3.2 Classification Analysis

For machine learning classification analysis builds a mathematical model through the identification of any given set of categories a new observation belongs, this when training set of data that contains instances or observations from whose category memberships are known [99]. The difference between clustering and classification in terms of machine learning is that classification is considered as supervised learning (learn from a training set of correctly identified observations), and clustering corresponds to the unsupervised learning that basically groups data into categories based on criteria of inherent similarities [100].

Since classification models are widely used for predicting and identifying customers' opinion, choices, and preferences based on previous events (historical data), we decided to explore these capabilities for meeting and addressing individual needs when building a product design. The next subsections show the considered algorithms or approaches for addressing this problem and complete the data analysis.

Decision trees

Classification decision trees are typically used for applying inductive learning algorithms to a set of training examples E . Each training sample $x \in E$ is a tuple $x = \langle x_1 = v_1 \dots x_n = u_n, r = l \rangle$, consisting of n feature value pairs plus a mapping for the single response variable r to a class label $l \in L$. For any new sample, $x' \notin E$, the classification tree provides a mapping $l \leftarrow f(x')$ [101].

Classification trees are constructed using recursive partitioning of the training dataset [100]. Algorithms in data mining use several splitting heuristics to estimate which variable x_i in E best explains the variation in the assigned values of r . Training samples then are split into 2 subsets (binary classification is assumed) so that the homogeneity of each subset concerning r is maximized. Each node leads to a path, and then defines a rule that consists of a conjunction of feature value pairs along the path. If the ancestors of i, a_i are defined as all the featured value pairs between the root node and node i inclusive, then the conditional probability distribution of r at i can be written as $p(r = l|a_i)$ [100].

Having the classification tree constructed, the classifier then needs to find the path through the tree that satisfies the feature value pairs in the unclassified example x' . The class label of x' is determined by the distribution of training examples at leaf i . Specifically, for the case presented in equation 9,

$$l = \arg \max_{l \in L} [p(r = l|a_i)] \quad (9)$$

Similar to decision trees, classification trees have the potential to grow exponentially large. The maximum number of leaf nodes in a classification tree based on n binary-valued attributes is 2^n . Nonetheless, unlike decision trees, the worst-case size is seldom realized for classification trees. Most inductive learning algorithms include features such as significance thresholds for splitting, a minimum sample size of leaf nodes, and validation pruning that result in parsimonious classification trees. The extent to which a final classification tree is smaller than the worst case is difficult to assess a priori. Moreover, tree size is determined to a large extent by characteristics of the training data, such as signal-to-noise ratio and interdependencies between features.

For this specific work, ensemble bagged decision trees were used for classification. Ensemble methods as described in [102, 103], are machine learning techniques that combine several learners, these include boosting, bagging, and stacking. These ensemble methods are often used to improve the predictive performance [104]. The difference between a normal classification decision trees and bagged trees is the combination of other learners since the model is constructed from multiple predictors using different training sets. Then the predictors are added from these models according to endogenously determined weights. Bagged trees use single base learner and choose random training sets combining the results of many decision trees, this reduces the effects of overfitting and model generalization is improved [104].

Support Vector Machine

The SVM is a well-known machine learning approach based on statistical learning theory [105]. The use of this classification algorithm for this work was based on its risk minimization and performance on learning tasks; SVM does not require prior knowledge, and a general behaviour description is guaranteed.

SVM's classification capability is based on the kernel function and penalty parameters. The goal of this linear-based classifier is to find the optimum decision region to get better generalizability with limited training data. Here, the boundaries or limits are set by the learning capacity of the machine. Mainly, SVM constructs an optimal hyperplane or maximal margin hyperplane as a decision surface in a way that the margin of separation between 2 classes is maximized[105].

The equation of separating the hyperplane is given by $(w \cdot x) + b = 0$, where w represents the vector of coefficients and b is a constant. The set of the 2 tuple training samples consists of the data vector x_i and its target or class y_i is included in equation 10 [106]:

$$(x_i, y_i), i = 1, 2, \dots, l, x \in R^d. \quad (9)$$

All the considered parameters should satisfy the following relationship, such parameters are represented for handling non-separable data:

$$y_i[(x_i, y_i) + b] - 1 \geq 0, i = 1, 2, \dots, l. \quad (10)$$

The class margin $p = 2/\|w\|$ reaches maximum through minimizing $\|w\|^2$. The following equation shows how the problem can be addressed:

$$\begin{aligned} & \min_{w, b} \frac{1}{2} \|w\|^2, \\ & s. t. \ y_i((w \cdot x_i) + b) \geq 1, i = 1, \dots, l. \end{aligned} \quad (11)$$

Usually, this is solved by giving a solution to the following problem:

$$\begin{aligned} & \max_{\alpha} -\frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l y_i y_j (x_i \cdot x_j) \alpha_i \alpha_j + \sum_{j=1}^l \alpha_j, \\ & s. t. \sum_{i=1}^l y_i \alpha_i = 0, \\ & \alpha_i \geq 0, i = 1, \dots, l. \end{aligned} \quad (12)$$

Thus, it is possible to obtain optimal solutions of the Lagrange dual problem since $a^* = (\alpha_1^*, \dots, \alpha_l^*)^T$. Worth to mention that forming the Lagrangian that involves constraints of the form $f_i > 0$, the inequality constraints equations are then multiplied by the nonnegative Lagrange multipliers (i.e., $\alpha_i > 0$), and then subtracted from the objective function. Then, it is possible to obtain the optimal $\alpha_i \geq 0, i = 1, \dots, l$.

$$w^* = \sum_{i=1}^l \alpha_i^* y_i x_i \quad (13)$$

$$b^* = y_j - \sum_{i=1}^l \alpha_i^* y_i (x_i, x_j) \quad (14)$$

The objective function can be given by solving equation (14).

$$f(x) = \text{sgn} \left(\sum_{i=1}^l \alpha_i^* y_i (x_i, x) + b^* \right) \quad (15)$$

The equation presented in (15) is the result of solving equation (14). Under the condition of linear inseparable and nonlinear, relaxation ε_i and kernel function $K(x, x') = (\Phi(x) \cdot \Phi(x'))$ are added to the inequality to solve the classification problem [105].

The next section discusses the feature selection approach used in this analysis.

3.3.3 Feature Selection Analysis

In machine learning, feature selection involves finding a subset of input features that best describe the underlying system structure better than all available features [107]. This approach complements well with classification since, without selecting the most significant predictors for the model, insignificant features might become noise and alter the performance of the model, therefore producing a not desirable result [108]. Although there are many methods for solving the feature selection problem such as incremental learning, neural networks, self-organizing maps, classification trees, fuzzy clustering, and GAs [109]; in this work we used GAs as the main objective function for selecting the best design attributes because of how powerful and convenient this approach is under certain conditions.

Feature selection is divided into two categories shown below.

- Wrapper methods, these methods use the output of the learning machine as selection criteria. On each iteration/step, the selected subset improves the performance of the previous one [110]
- Filter methods, a faster convergence is obtained because it is used as an indirect measure of the quality of the selected features. The right subset might fail to be selected if the criteria used is diverted from the one used for training the learning machine [111].

In both cases, GAs had been used to solve feature selection problems obtaining good results and performance of the classifier [108].

When feature selection is considered as a learner, from the sample scheme it can be described as: given a set of labelled data points $\{(x_1, y_1), \dots, (x_l, y_l)\}$, where $x_i \in R^n$ and $y_i \in \{\pm 1\}$, choose a subset of m feature ($m < n$), in which the lowest classification error is achieved. In [112] feature selection is defined as finding the optimum $n - column$ vector σ , where $\sigma_i \in \{1, 0\}$, that defines the subset of selected features, as found as:

$$\sigma^o = \arg \min_{\sigma, \alpha} \left(\int V(y, f(x * \sigma, \alpha)) dP(x, y) \right). \quad (16)$$

Where $V(\cdot, \cdot)$ is a cost function that maps the values, $P(x, y)$ is the unknown probability function from where the data was sampled and is defined $x * \sigma = (x_1 \sigma_1, \dots, x_n \sigma_n)$. The $y = f(x, \alpha)$ function, is the classification engine that is evaluated for each subset selection σ and for each set of its hyper-parameters α .

The objective of this approach is to process the data in order to extract, potentially useful, novel, valid, and understandable structure in data by identifying relevant and meaningless features [113].

From the other hand, what GAs represent for feature selection, this approach represents a type of robust problem-solver based on a population of solutions that evolve through consecutive generations by means of the applications of three genetic operators: mutation, crossover, and selection [114]. This approach is suitable when performing exploration in huge search spaces, where other methods (gradient, or local searchers) cannot find good results. In the case of feature selection, it uses an encoded binary representation of the chromosomes from which then the evolution of individuals starts to take place. In Figure 3-4 is presented the basic steps for feature selection considering genetic search.

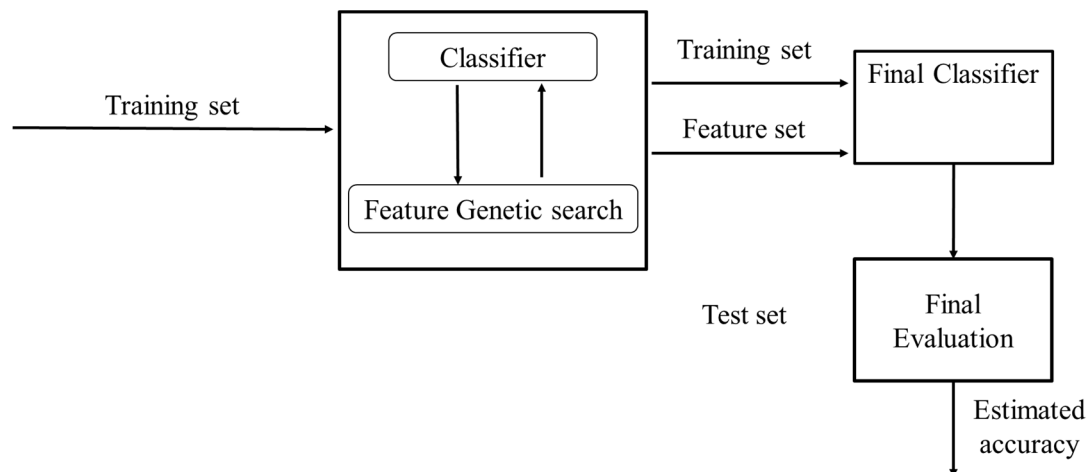


Figure 3-4 Framework for the Feature Genetic Search [108].

In the next section, we present how all these approaches and algorithms come together in this work, as well as how we decide to use the machine learning tool discussed previously. In the next section is presented the perspective of design for smart environments and i4 principles.

3.4 Smart Design Under Industry 4.0 Principles

Technologies inside a smart factory can also enable communication to inform a virtual copy of the process to personalize designs before actually processing in the physical

world. The importance of having a good product design can be decisive in terms of quality, performance, and customer acceptance. The end-user experiences when a product was designed according to their needs and wants, but this is never an easy task because this involves putting together the right materials, the proper technology, and the adequate hardware aiming to bring the best experience to customers. In the era of Internet and information and communications technology, i4 concept lead to a new perspective: adapting to individual requirements by bringing flexibility to manufacturing processes with the capabilities that CPS and IoT give. The key challenge for manufacturers and designers is to understand how to harness and use knowledge to innovate goods and interaction with customers.

Using all these innovative ways of producing a design that considers characteristics like focused, informed, and refined according to individual needs and wants is what we call a “smart design” [115]. Is implicit in smart technologies that the system tries to give each user the opportunity to have a personalized experience, this is why in this work we focused on i4 principles to achieve customization. Moving forward with these concepts and approaches for customizing products according to smart design principles, the existing technologies for personalizing product design need to be discussed. The role of automated design for i4 is discussed in the next subsection.

3.4.1 Automated Design for Industry 4.0

Used for smart manufacturing environments, the automated design in this sense is used in this section as CAutoD, and it aims to reverse a design problem to a simulation problem, then automates such digital prototyping by an intelligent search using biologically-inspired machine learning, hence accelerating and optimizing a human trial-and-error process in the computer prior to physical prototyping. As discussed before concepts or tendencies like Industry 4.0 has to integrate several frameworks, the main tool for CAutoD is evolutionary computing, including GA, particle swarm optimization (PSO) and ant colony optimization (ACO). Intelligent system utilises such

computational intelligence to analyse interactions between variables or phenomena, so as to identify causes, effects, drivers and dynamics for their modelling, design and control in a holistic manner. Since the purpose of this work is to match concepts where i4 and networked production meet, biologically-inspired evolutionary computation used for search in multi-objective designs, for optimisation of system structures (as well as their parameters), and for intelligent and automated virtual prototyping.

According to [26] a design problem is concerned with finding the best parameters within a known or given range through parametric optimisation or learning and is also concerned with inventing a new structure beyond existing designs through structural creation or machine-invention. If the objective cost function $J \in [0, \infty]$ (or, inversely, the fitness function $f = 1/(1 + J) \in [0, 1]$) is differentiable under practical design constraints, the problem is solved analytically. Then the author points that unfortunately, this scenario does not usually exist in practice and the problem is hence often unsolvable, since the cost function of J is minimize (the lower the better) and for f (fitness) the cost function is maximize, in practice the derivative is difficult to obtain and many peaks will be encountered. In Figure 3-5 it is shown the evolutionary computing transforming process, where the research focus on control systems, like Computer-Aided Control System Design (CACSD) and Computer-Automated Control System Design (CAutoCSD) and how it is transformed manually.

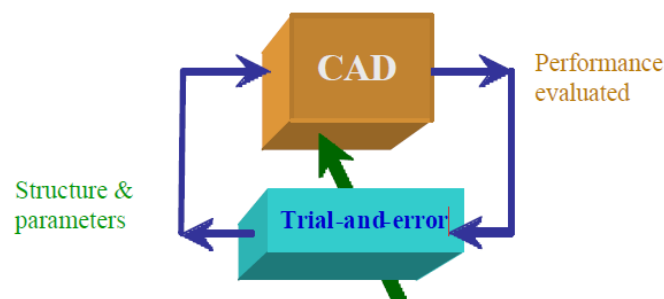


Figure 3-5 CAutoD realised through an evolutionary computing process [26]

The key of this approaches to work is to reach the evolved top-performing candidate prototypes will present multiple optimal designs and the Evolutionary Algorithm based on CAutoD can start from the designer's existing database or even randomly generated candidates.

Future directions at the moment point that the trend of Intelligent CAutoD for i4 may have seamless CPS integration to deal globally with:

- Predictive data analytics to extract emerging trends in societal needs and wants then enhance conceptual designs for smart manufacture.
- Transform digital prototyping (CAD) to automatic and optimal virtual prototyping (CAutoD) on the cloud
- Reduce traditional product development cycle from:

concept → prototype → test → fix → manufacture

To:

concept → design, innovate or create → manufacture

Many researchers suggest that experimental research should be considered in order to validate scientific results of the theoretical work. Validation and implementation of these approaches will help with a fast rhythm of acquiring knowledge and developments. Meanwhile, for today's perspective of i4 is still in process of implementation, there is still a way to help in a design and manufacture perspective, this can be through the adaption of concepts like CAutoD [7].

Knowing that most of the research and future developments point to those directions already discussed, the focusing when talking about i4 is the Integration of a well-funded methodology for CPS, available technology and infrastructure, intelligent approaches that allow automation as well, and as well as considering that everything goes through the IoT and IoS. There are enough tools nowadays that can be used to develop analysis, which therefore is another focusing that most of the work discussed highlights.

3.4.2 Motivation of Selected approaches

The selection of algorithms for classification and feature selection has been determined based on the literature review on the state-of-art classification tools, discussed by [66] and [99]. For this comparison of approaches, it is necessary to consider the following questions to assess the output models:

- How much detail is required?
- What type of data is used to build the model?
- How much data is provided, and is it continuous?
- How important is it to visualize the process?
- What do we want to achieve?
- Is storage a limiting factor?
- Are the considered inputs numerical or categorical?

The prediction of customers' needs and wants is central in this case. The methodologies that will help to realize prediction of attributes for smart design under i4 principles are as important as the prediction of customers' needs and wants. Here, for i4 environments, authors like [3],[16], and [49] highlighted the importance of considering approaches that help to visualize the problem and help human experts involved in the processes of decision-making and understanding the data move in the right direction. This is the reason why approaches like decision trees were considered for testing the cases of study.

Since the scope of this work was to address customer needs and wants using the principles of i4 and smart design, it was found in [4, 81, 91] starting points to consider how to mine and analyse the data, in order to predict what customer needs and wants can possibly be. The literature review helps us to explore the capabilities that machine learning bring when dealing with complex problems, pattern recognition, data classification and meet individual needs predictively [27]. In this sense, dealing with mass customization can be a very complex task to tackle; however, feature selection using GAs can help narrow down options, and act along with classification as a recommendation method when customers need to select the most suitable design. Once the selection of attributes accurately describe customers' needs and wants, product design can constantly improve and the system can make better prediction thanks to digitalization [4].

The proposed framework involved the following steps:

- Data preparation;
- Selection of an algorithm;
- Fitting the model;

- Choosing a validation method;
- Examining fit and updating until satisfied;
- Using a fitted model for prediction.

Preparing the data help to determine which attributes can be used as predictors and which can be used to understand the problem, making this stage a crucial part. According to [66], all supervised learning methods start with an input data matrix, usually called X in this case. Each row of X represents 1 observation. Each column of X represents 1 variable or predictor. Then, the missing entries are represented with not-a-number values in X . For each dataset, it must be determined whether a variable is considered a predictor or a response. Some variables must be disregarded or not considered inside the prediction model since they are not significant for the description of the problem.

In the context of this work, the focusing of predictive analytics used to address customers' needs and wants match perfectly with that used in business intelligence. In business intelligence, predictive models extract patterns found in historical data, focusing on identifying opportunities and risk [51]. Predictive analytics provides a predictive score for each value with the purpose to inform, determine, or influence organizational processes that belong to a large number of individuals (customers) [116]. The types of models that encompass predictive modelling include the following [59]:

- Predictive models: Models obtained from the relation between a specific performance of a unit in a sample and 1 or more known characteristics (attributes) of the unit. These models aim to evaluate the probability that a similar unit in a different sample will exhibit a specific performance [59]. In this work, predictive models in this sense correspond to pattern

identification, aimed at identifying and predicting customers' needs and wants.

- **Descriptive models:** Here the relationships in data are quantified in a way that is commonly used to classify prospects or customers into groups. Descriptive models identify different relationships between customers or products. In these models, a rank-order of customers is not found. The probability of taking a particular action in the way that the predictive models do is also not found [59]. Descriptive models can instead be used to categorize customers by preferences on products, which reflects the main objectives of this work.
- **Decision models:** This type of models describes the relationship between all elements of a decision. Its purpose is to predict the results of decisions that involve several variables. Applications of decision models include optimization and maximizing certain outcomes while minimizing others. Matching applications of decision models to what is proposed in this work, we developed decision logic rules (business rules) to reflect the desired actions for individual customers or circumstances [59].

Although there are numerous projects and researchers using AI, machine learning, and digital models to address customer satisfaction, there is not enough evidence to support an effective integration or a methodology that encompass smart design, mass customization, and prediction of customer needs and wants using i4 principles. As discussed before, many companies address differently the customization problem, the solutions are very diverse. Literature review spot out a common idea between manufacturers, which is how to make production stage more effective in terms of costs, complexity, and time; but not many study cases focus really on the i4 perspective— what customers' needs and wants really are [117].

3.4.3 Case Studies of Predicting Potential Customer Needs and Wants for Future CAutoD

Industry 4.0 and big data studies have led to our AI-based framework of a closed-loop value chain for smart manufacturing with CAutoD as an engine for smart design. To apply the methodology of attribute prediction using smart design principles, this section analyses 8 applicable cases.

Use case 1: Smart Remote Machinery maintenance [16]. Application development for a heavy-duty equipment utilized in mining construction, the author includes several health prediction tools for a diesel engine component. In this case an application for remote monitoring, data is acquired on a daily basis that includes parameters from the diesel engine to the remote location, those parameters include: fuel flow rate, pressures, rotational speed of the engine and temperatures. For the output, it was necessary to assess the health of the engine, determine what causes the abnormal behaviour, then predict remaining life of diesel engine. Using a virtual suite called Watchdog Agent® toolbox, allows converting the engine data into health information. Including Bayesian Belief Network (BNN) to classify different engine patterns in the data to build a model, and this makes suitable the interpretation of anomalous behaviour, therefore detect a problem with the engine on an early stage of degradation. Finally, for prediction it was used by [16] a fuzzy logic-based algorithm or fuzzy membership, minimizing uncertainties in data and making more robust the approach.

Use case 2: Kaiserslautern Smart factory project [118] (source: Siemens / Bosch) Part of the German Centre for AI (DFKI) which demonstrates by the use of soap bottles indicating how assembly lines and items can communicate each other. Empty soap bottles have labels with RFID, those labels communicate with machines and inform if the jar must give a white or dark top. With radio signs, an item is constantly speaking with its surroundings and transmits an advanced item memory from an earliest starting

point. For this case, CPS empowers reality and virtual part of the process which are constantly combined.

Use case 3: Goal control 4D [119] (source: FIFA/ Goal Control GmbH). By integrating 7 cameras per football-goal entry and located on the rooftop around the football field, cameras are associated with a high-performance PC, tracking the development and movement of all individuals in the field (players, officials, and minor elements). The football is the most vital individual, the position is constantly followed and retrieves three coordinates or measurements: X, Y and Z which measures with an accuracy of millimetres every time the ball gets closer to the goal line. If the ball crosses the goal line, in one second the Central Evaluation Unit sends an acoustic and optical sign-in to the collector clock of the mediator. Instantly the cameras record the pictures of the event, in order to accept the goal. The word: “Goal” appears on a watch that the official has on his/her hand. Other companies are running some tests to use this same technology in the automotive industry by supporting virtual accident tests, minimizing expenses of raw material, test-hours and time.

Use case 4: High-end centralized computing for Husky [120] (source: Beckhoff/Intel) Collaboration between Beckhoff and Intel for developing processors and Information Processing Centre (IPC) for Husky company which is based on Injection molding systems that manufactures equipment used in a large range of plastic products (closures for beverages, bottles and parts for the medical industry). Part of the challenge was the achievement of system accuracy, responsive machinery dynamics, and repeatability when designing injection moulding machines; then as well system design approach, minimize total cost to produce, at the same time ensure high-quality performance. Committed to accelerate Industry 4.0 development, Intel and Beckhoff point out the following technologies that suit best the challenge presented by Husky: ➔Intel IoT Developer kit (variety of programming environment, tools, hardware, application programming interface, and cloud connectivity solutions); ➔low-cost

Open Platform Communication (OPC) servers for communication to MES; → Intel IoT Gateway Development kit used to create a fast prototype that is reliable and scalable providing communications, security, manageability among other functionalities; → IPC with Intel multicore processors capable of energy management, condition monitoring, and highly integrated machine designs (integrate robotics); → The Windows Control and Automation Technology (TwinCAT) with Matlab interface for creation of process simulation environments for virtual commissioning; → Automation interface in TwinCAT for remote access to control programs and to dynamically change those based on production situation. Encountered results show the launched injection moulding system called: “HyPET* HPP5” equipped with Intel Core i7 and high-end computing power by Beckhoff C6930, which provides productivity and cycle gains from 3% up to 12%.

Use case 5: Self-organizing adaptive logistics (source: Daimler) [118]. Here it takes place the Product Life Cycle and lifetime. Inside networked production, reliability for production logistics processes is crucial for friction- and error-free production processes. Automotive industry requires adaptability, amount, variety and option accessibility of required parts and supplies. CPS allows transparency in material and logistic parts. Integration of CPS allows material and development of parts to optimize the entire supply chain. It serves as the technical foundation for a dynamic intra-logistic controlling in flexible factories.

Use case 6: Customer integrated engineering [118] (source: IPA). The ever further-reaching client requirements, adherence to deadlines and late changes are driving the necessity for a fundamental shift within the interaction between classical production tasks and the customer or the supply chain. Integrating consumers in the developing, planning and value-added activities of the shrunken company results in novel transparency and a reactive production in perfect synchronization with all the customers involved.

Use case 7: Production line for composite components with a gripper spider of 15 needle grippers [118] (source: SCHMALZ) Used in textile industry, it is required a flexible production of fibre composite components, each product changeovers per day, each product requires for different material thickness distinct needle stroke. The solution proposed by SCHMALZ includes the needle gripper SNG-AE for handling highly porous and non-rigid materials. For each cycle, the stroke can be adapted in any order, as well as stacking with the use of one single gripper and bidirectional interface for enabling communication between the higher-level field-bus systems. This case represents a higher benefit for customers because of the elimination of downtime during production changeovers, setup time, increased flexibility of production, minimal risk, error correction in planned maintenance and error detection during operation.

Use case 8: Smart factory architecture [118] (source: IPA) Along with the thought of a product's lifecycle, several companies have already begun thinking about the factory's lifecycle. It is remarkable how difficult synchronizing these lifecycles actually is. Analogous to those lifecycles, a smart factory has its own lifecycle that can be designed in accordance with the product. The smart factory offers an opportunity to establish a comprehensive lifecycle by associating an HTO approach with IT on a meta level.

Collaborations to trigger necessary technology for Industry 4.0 are found in many study cases included in this section, where making the system interactive with human support is key. By keeping a simple and useful interaction of virtual-physical-human part of each process is essential, visualization tools can represent for human tasks a high-value development, in order to constantly supervise the process and minimize performance errors.

3.5 Summary

As design attributes play a role of customer needs and wants, there is a way of informing the manufacturing process what attributes are more desirable for individual users. We shall, therefore, choose a design for manufacture data as four case studies in illustrating the methodology and its applications in this thesis. Considering this, we shall develop a framework that integrates the technologies and approaches discussed above. In the following chapter, we describe the framework for predicting customers' needs and wants.

Chapter 4 Framework for Predicting Potential Customer Needs and Wants

In this chapter, the frameworks that we propose are presented and discussed, in order to predict customer needs and wants. It is clear that the first step needs to include a closed-loop value chain framework that encompasses prediction of customer needs and wants, in a general way. Once this methodology is proposed, the second stage needs to cover specific steps inside the closed-loop framework that includes AI. Next stage narrows down the AI approaches to test classification models and train it from historical datasets and also pattern recognition with cluster analysis. In the end, the selection of best design attributes framework is proposed.

4.1 Value Chain for Predicting Potential Customer Needs and Wants

This framework is proposed as a first stage and after revising the literature and common research work. For this we decide to give an answer to the six questions stated in chapter 1, section 4, presented as follows:

1. Where in the industry value chain, most value is added?
2. What are the benefits of Industry 4.0 to the customer?
3. What are the major challenges in Industry 4.0?
4. How to design smart products efficiently?
5. How Industry 4.0 will add most value/most efficiently?
6. What benefits will Industry 4.0 bring to the manufacturer?

Taking these questions into account, the answers found are presented next:

1. Most value is added for customers, since they play a key role and for i4 principles, smarter way of manufacture products adds value, the potential lies on highly customized products at mass production costs. i4 allows for a faster response to customer needs than is possible today. It improves the flexibility, speed, productivity, and quality of the production process.
2. The benefits of i4 to the customer, is discussed in [31, 121] that the main benefits are mass customisation, opportunity for self-designed and locally-made unique products, but also be a chance for new business models, and as an example companies like YouTailor®, Bombsheller® and MyMuesli® are offering through their website products that cannot be found in the store shelves, demonstrating that this is not a vision of the future, beyond that, is a necessity from the customers.
3. The major challenges are the integration of technologies and drivers for i4. These technologies need to be part of a methodology that effectively and intuitively can also address a high level of product variety at mass production costs, and at the same time fulfil individual desires.
4. Designs are improved from a digital platform that considers the analysis obtained from historical data. CAutoD is the approach that best matches with this problem of design since involves AI and automation.
5. Value occurs when needs are met through the provision of products, resources, or services; and finally, the value is an experience and it flows from the person (or institution) that is the recipient of resources, it flows from the customer. These concepts point to what is a key difference between a value chain and a supply chain, they flow in opposite directions [33].

6. The benefits that manufacturers get from i4 include (i) operational costs reduction thanks to the interlinked devices through a network and embedded computing, (ii) productivity increment thanks to the flexibility, more efficient processes, and improve the decision-making process; and (iii) customer satisfaction increment thanks to digital systems able to tune product designs to meet customer needs.

Using this information for proposing a methodology to effectively integrate the concepts revised and the encountered challenges for i4, we decided to focus on a single task, which is to achieve customer satisfaction in a predictive way. The importance of having embedded in the product design process a virtual copy that can be modified and tuned according to customer needs and wants is key for addressing customer satisfaction. Thanks to the literature review, it was clear for us that the CAutoD principle matches perfectly with the design process since involves AI and most important, gives the opportunity of keep improving or evolving better designs through automation. In Figure 4-1 is presented the proposed framework that includes value chain and supply chain for predicting customer needs and wants in a closed loop.

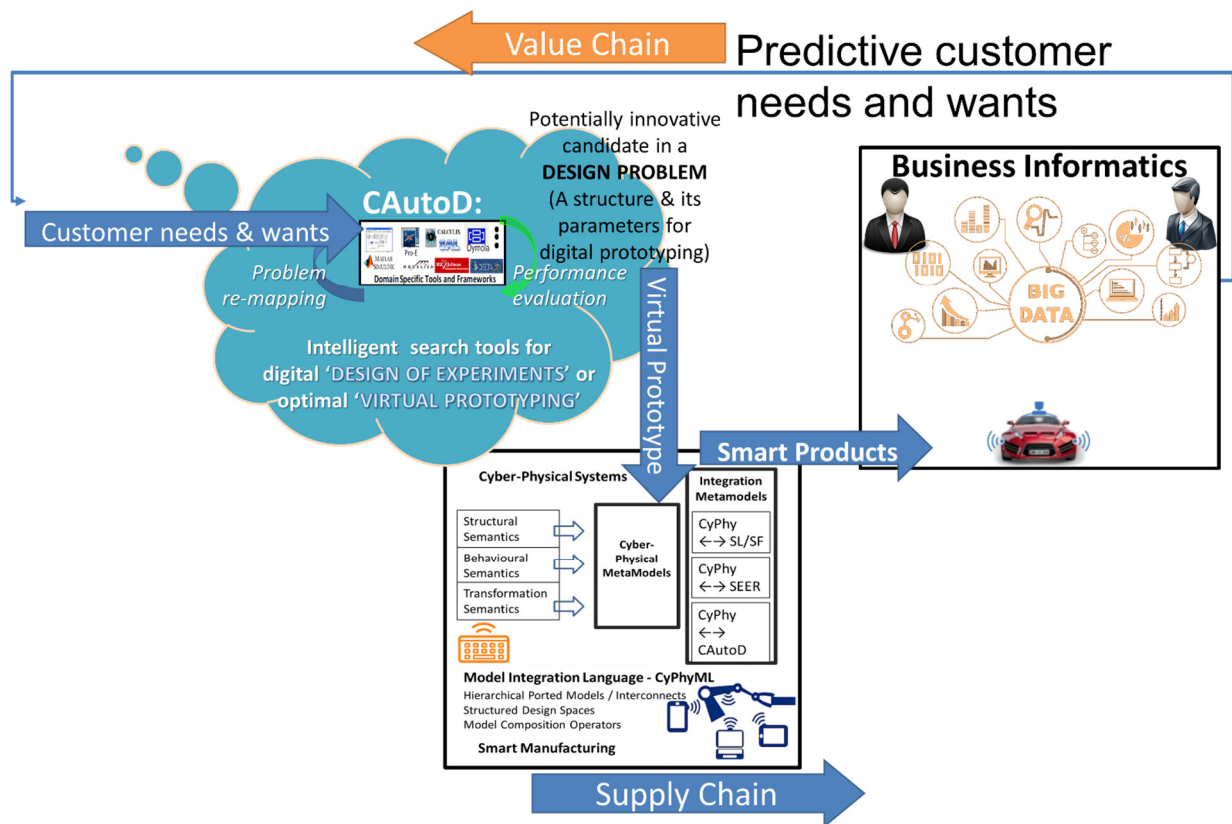


Figure 4-1 Value chain closed loop for predictive customer needs and wants.

With this methodology, a full integration of i4 components, revised from the state-of-the-art research, is included. For us is clear that for having a predictive feedback for addressing customers' needs and wants it necessary to integrate all these technologies, which lead to the following partial conclusions:

- Design process is suitable for automation with CAutoD
- Intelligent search within the design process allows needs and wants to be covered if the correct data is fed-back.
- CPS interconnected to the virtual prototype obtain the optimized design and manufacture it.

- The smart product is obtained and is business informatics role to obtain the necessary and reliable data that is going to be fed again into the loop.
- Since everything is connected through the cloud this enables to make it fast
- Decisions for manufacturers are easier to make, with automation.

Being this the first stage of this research, and a starting point, we decided to focus on the AI inside the closed loop and use the CAutoD principle to put in practice the capabilities that these approaches can bring. In the next section is presented the AI process for addressing customer needs and wants.

4.2 Artificial Intelligence for a Closed-Loop Framework

Figure 4-2 depicts the framework proposed to solve several of the aforementioned challenges in i4. Based on i4 and smart manufacturing's key objective, i.e., achieving self-prediction, being self-configurable to manufacture products, and providing services tailor-made at mass production rates, we propose a closed-loop framework that integrates several approaches from AI, concepts from smart environments (Smart Services, Smart Design, Smart Products, and Smart Manufacture), and the IoT feature/connection to analyse big data on cloud services [27].

This framework is presented in [27]. In the first block from Figure 4-2, customer needs and wants are first captured and processed to extract key design characteristics, here is also where the inputs taken from data are first encountered. This block is similar to what is presented in Figure 4-1. This information (collected inputs) is then fed into a CAutoD engine [26], where the design requirements, features and performance objectives are mapped into 'genotypes' for further analyses. This process, which is commonly known as rapid virtual prototyping, uses intelligent search algorithms such as the GA or particle swarm optimization to explore the design search space for

optimal solutions. In the proposed framework, this process takes place over the cloud and produces a set of optimized virtual prototype at the end of the search.

The second block of the closed loop in Figure 4-2 shows the virtual prototype, which is obtained from the selection and design process in CAutoD. Through the integration of CPS or cyber-physical integration, the virtual prototype in the second block is transformed into a physical product, i.e., the smart product. This block is where the smart product is manufactured using intelligent approaches.

The next part of the framework refers to business informatics and how the smart products are connected to the IoT. Here is where big data takes part. Through product performance and feedback from the customer, more features can be considered. This covers the necessary attributes that make the product manufactured in an optimal way.

Following this, the response obtained from the customer is automatically fed back to the system for further analysis and to fine-tune the virtual prototype. This part of the closed-loop cycle can be considered as the validation of customer needs and wants, where when necessary such validation can lead to better designs and upgrades to the current one, here inputs can also be fed-back into the smart design block. This analysis uses node or dynamic analysis that can perform clustering, selection, and detection of patterns, and visualization. After that, the fuzzy c-means clustering completes the update of selected attributes by comparing the latest input to the existing cluster and tries to identify the cluster that is most similar to the input sample. Then, several features are fed back into the cloud again. This process takes no more than 5 minutes to complete in theory, in practice really depends on the nature of the problem, as an example in [122] hull designs involves designing process of weeks, and to put together the stakeholders for adjusting, customising, and making

any alteration to the actual ship or hull is very difficult, retrieving useful information to have a smart design can take short time, but the manufacturing cycle takes months.

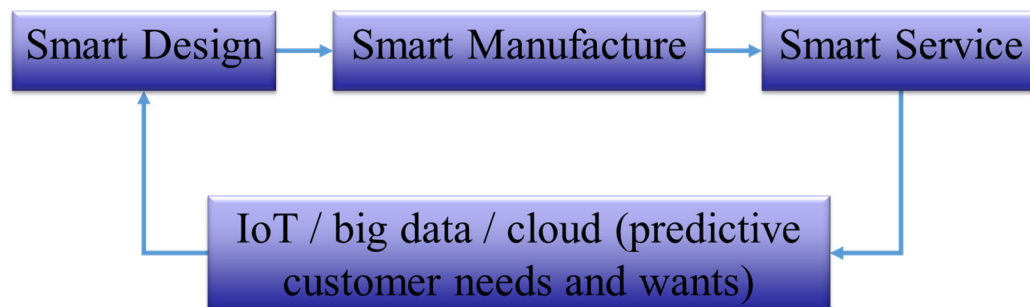


Figure 4-2 Industry 4.0 automated closed-loop for predicting customer needs and wants for customization [27]

The analysis can result in 2 outcomes [27]: (i) Similar clusters are found. This will be reflected as an existing attribute, and the algorithm will update the existing cluster using information from the latest sample. (ii) Non-similar clusters are found. The algorithm will hold its operation with the current sample until it sees enough count-of-cluster samples.

When the number of out-of-cluster samples exceeds a certain amount, there exists a new behaviour in the data that has not been modelled. Then, the algorithm will create a new cluster to represent such new behaviour. For these cases, feature selection approaches can be very adaptive to new conditions. Next, self-growing clusters were used as the knowledge base for customization assessment.

The next section presents an additional step used to classify the inputs and build the predictive model, this process takes place inside the previous framework presented in Figure 4-2.

4.3 Classification Learner Framework for Coding Customer Needs

Inside the closed-loop cycle takes place another process that involves data classification as presented in Figure 4-2, data accessing, validation, and data analysis using AI. This process encompasses the data analysis in detail, and all the components are presented in Figure 4-3. Inside the proposed framework, the prediction models, classification of attributes using the machine learning approaches for classification and clustering analysis are obtained.

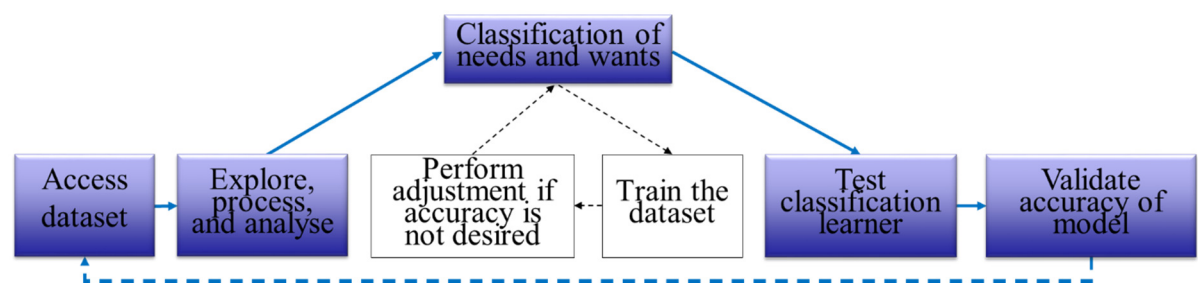


Figure 4-3 Proposed AI-based methodology for predictive data analysis and attribute classification.

The link between Figure 4-2 and Figure 4-3 is the process of predictive customer needs and wants after receiving feedback from the smart service, this creates a constant feedback that creates a more informed model, therefore a more robust system. In Figure 4-2, the block that corresponds to the predictive customers' needs and wants connects to the IoT. Using big data analytics is unfolded into detailed steps in Figure 4-3. The processes of accessing the data, exploring the data, developing the model, test classification, and validation of classified attributes take place inside the closed-loop in Figure 4-3. The significance of this also relies on automating the process of tuning product designs using the classification model, this allows customizing designs digitally before the manufacturing process begins.

This is achieved using self-organizing maps against fuzzy k-means, bagged decision trees, and support vector machine approaches. Hence, this work built better and less complex models to help visualize patterns, determine interactions between variables, and predict and select attributes. This is useful in the i4 value chain to address customization and improve the decision-making process. One of the main discoveries using several learning approaches is that decision trees give a more accurate analysis and are easy to interpret, but this will be discussed in detail in the following chapters.

The next section presents the integration of feature selection approach to the closed-loop cycle framework. This completes the proposed methodologies for predicting customer needs and wants using i4 principles.

4.4 Genetic Search Framework for Selecting Best Attributes

To complete the whole closed-loop value chain for predicting customer needs and wants and suggest product design alteration to meet customer satisfaction we decided to add other steps to the framework. Figure 4-4 illustrates the framework proposed to solve several of the aforementioned challenges in i4. Based on i4 and smart manufacturing's key objective, i.e., achieving self-prediction, being self-configurable to manufacture products, and providing services tailor-made at mass production rates, we propose a closed-loop framework that integrates several approaches from AI, concepts from smart environments (Smart Services, Smart Design, Smart Products, and Smart Manufacture), and the IoT feature/connection to analyse big data on cloud services [27].

In the first block from Figure 4-4 and identified with number 1, customer needs and wants are first captured, described as the data acquisition where all the necessary information is gathered. In the second block, identified with number 2 is presented the mining of customer requirements, such requirements are processed to extract key design characteristics/elements. Inside this stage, it is proposed a CAutoD engine

[26], where the design requirements, features and performance objectives are mapped into 'genotypes' for further analyses and/or selection. This process, which is commonly known as rapid virtual prototyping, uses intelligent search algorithms such as the GA or particle swarm optimization to explore the design search space for optimal solutions. In the proposed framework, this process can take place over the cloud or data mining and produces a set of the optimized virtual prototype at the end of the search to be recommended to the user.

The third block identified with number 3 of the closed loop in Figure 4-4 shows the modelling part, which is obtained from the selection and design process in CAutoD and through mining customer requirements. Through the integration of CPS or cyber-physical integration, the virtual prototype in the third block can be transformed into a physical product, i.e., the smart product.

The next part of the framework identified with number 4, refers to the validation of the model. This part differs from the one inside blocks 2 and 3 because this validation represents the feature selection and classification models together, where in the previous blocks the classification and selection were performed and obtained. The obtained results from the trained dataset are validated against new inputs and check the corresponding accuracy level of the model to see if it is good for making predictions and proceed with following steps. This covers the necessary attributes that make the product manufactured in an optimal way.

Following this, the next block (5) will make a recommendation to the user based on the trained classification and feature selection models. The response obtained from the customer is automatically fed back to the system for further analysis and to fine-tune the virtual prototype. If the requirements of customers are not met, the trained dataset needs to be evaluated again, or analysed for better adjustment. This analysis uses node or dynamic analysis that can perform clustering, selection, detection of

patterns, and visualization. After that, the decision tree classification completes the update of selected attributes by comparing the latest input to the existing set and tries to identify the attribute that is most similar to the input sample. Then, several features are fed back into the cloud again.

Block number 6 corresponds to the part of automation and control. It is suggested in [26] that constant development of models can result in a time-consuming task, but once your tested model gives accurate results, adaptive control can help to maintain the predictions and give a certain level of automation to maintain the process through time and constantly making predictions. In this specific case, feedback adaptive control can be useful based on the nature of the problem, i.e. closed feedback loop that retrieves information constantly. The adaptive control as the name suggests, it will adapt to the controlled system proposed in Figure 4-4, and more specifically to the customer needs and wants coded into design elements, trying to make an iterative learning control system to constantly found the best design attributes.

At the end of the closed loop cycle presented in Figure 4-4, the last block (7) represents the customer satisfaction fulfilment. Here the idea is to maintain a constant communication with the user of the product and being able to measure how satisfied an individual is with the recommendations and selections made by the system. In this part IoT and cloud computing are used to improve predictions of the system, and different to what was discussed about stage 4, the validation does not happen internally, this validation using data services and cloud is an external validation or a real indicator of addressing customer needs and wants.

The above-presented framework completes the full value chain methodology shown in the first section of this chapter. This framework was proposed as a result of exploring the capabilities of data mining techniques. At the beginning, suggested by literature review, the framework seemed to be a matter of integrating technologies

and AI approaches for customizing product designs using big data analytics. In practice, we explored a more refined AI-based method for predicting or recommending to customers what attributes are most likely to be selected from a wide variety of options, since in Chapter 2, section 5 it was discussed that customers sometimes are not clear what their needs and wants are and put in practice the benefits of automation to make the decision-making process easy to both, customers and stakeholders. The transition of methodologies from Figure 4-1 to Figure 4-4 is due to experimenting with different case studies that represented different challenges on how to analyse the data, but still trying to obtain similar results.

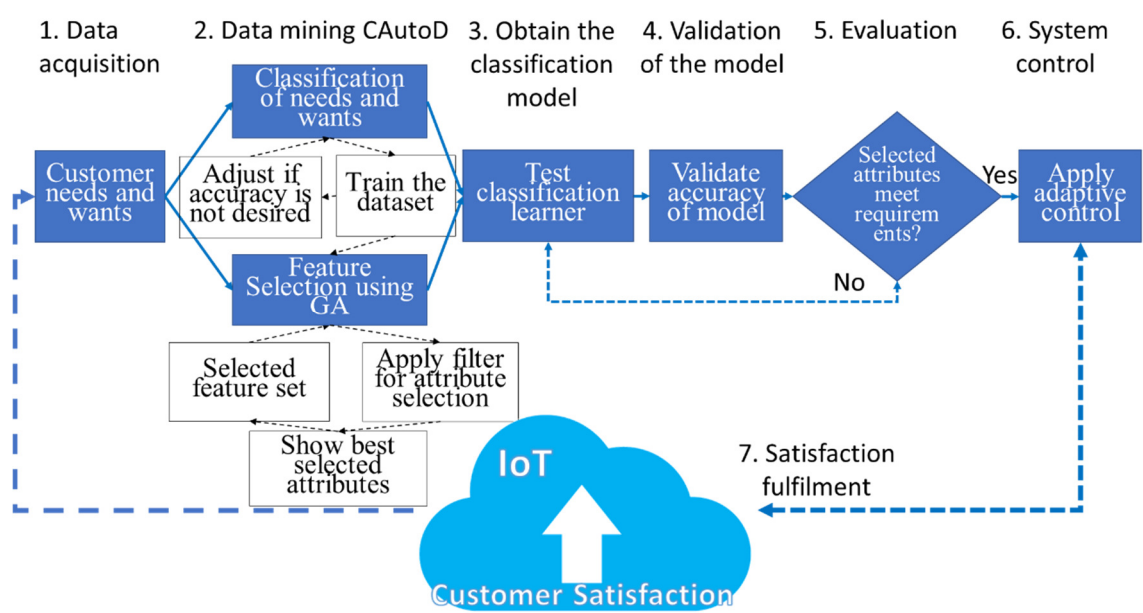


Figure 4-4 Industry 4.0 closed-loop for predicting customer needs and wants using data mining approaches

Inside the final closed-loop framework presented in Figure 4-4, there is another methodology that represents only the process involved for feature selection and genetic search. Figure 4-5 depicts the used methodology inside the full closed-loop framework presented before, the difference is that this framework combines the GA steps with the feature selection for mining the design attributes. This framework

describes the common approach used for feature selection, combined with the well-known approach for GA found in [114]. The common procedure for feature selection is described in chapter 3, section 2, and subsection 3 (3.2.3), the difference here is that the actual process inside Matlab program is depicted here, where we interlinked the obtained results in feature selection with the machine learning toolbox for obtaining a more accurate result. As described before, this process also involves feedback from the classification modelling process, once obtained the training dataset and is used to perform the feature selection. The objective function used in this step, nested inside block 2 of Figure 4-4, corresponding to the lower part (feature selection), was targeting the design attributes that appeared in the classification analysis as most significant. This was obtained as described before, by using the trained dataset and performing a feature selection algorithm.

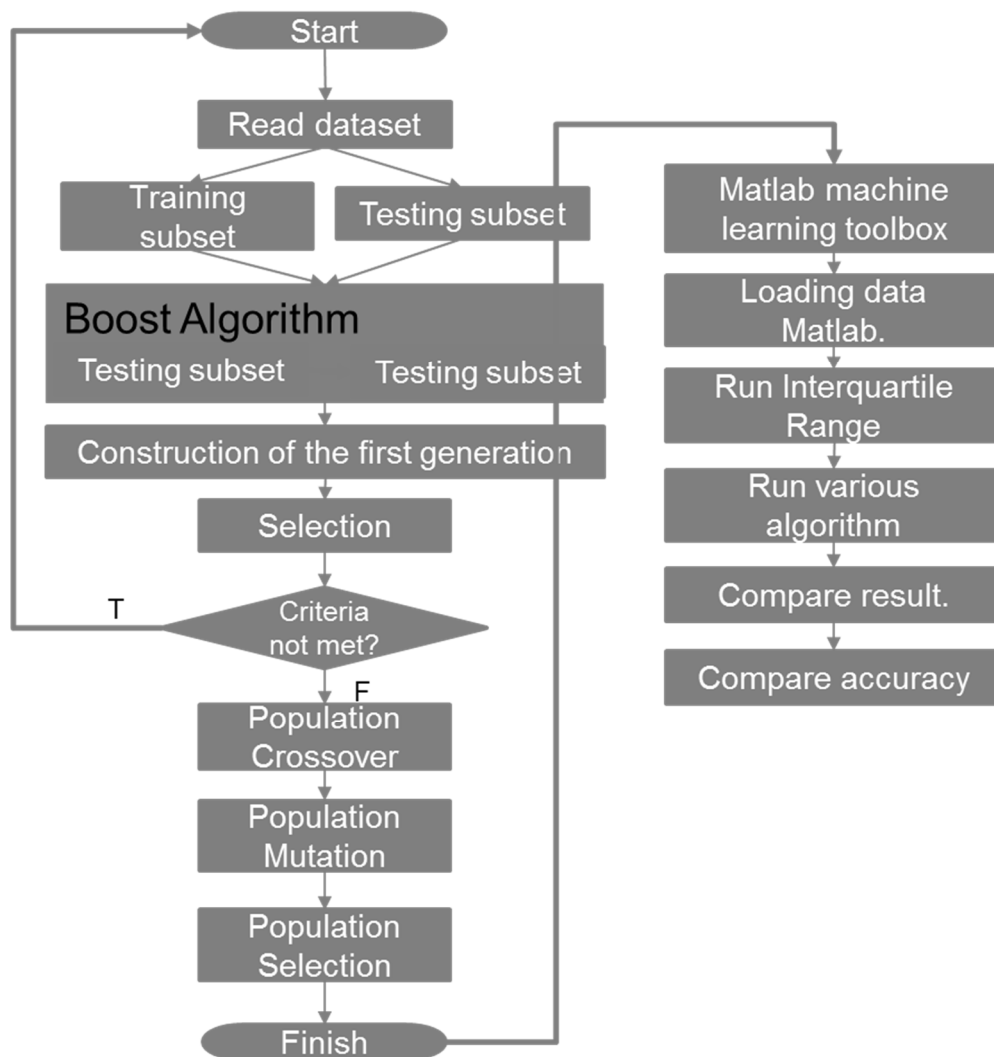


Figure 4-5 Genetic search framework for feature selection.

Finally, the next section gives a summary of the discoveries obtained when using this different methodologies and approaches to predict customer needs and wants or recommend a set of features to complete the product design tuning for addressing individual needs.

4.5 Summary

In this chapter all the proposed frameworks were presented, the objective behind was to obtain a predictive way to customize product designs that meet individual needs and wants. The first framework shown in Figure 4-1 has been obtained out of the literature review and used as a starting point. In this first framework, it was clear that a closed-loop was more likely to fill the gaps missing in the process of personalizing design products, and smart technologies were able to improve this process. An application of this framework is found in [122], where it was decided to adapt the approach to an energy-efficient manufacturing process for ships and vessels, acquiring the perspective of through-life smart design and operational process as well. A two-way closed loop that addressed the needs of a specific manufacturing process, in this case, ships and vessels represented a perfect example of considering a smart way of designing the products, because of the challenge, effort, and money that this market entails.

Moving forward with predicting customer desires, we decided to move from the value chain proposed framework to a simple framework that encompassed the key drivers and concepts involved in what we wanted to achieve, i.e. predicting customer needs and wants. Here a number of applications were used and published in [27, 62, 123], and all these case studies use classification methods for mining customer needs and wants. In practice and throughout all these applications was discovered that AI enables one of the key principles of i4 – self-adapt. The use of machine learning approaches to reach the vision of i4 and smart factory concepts now were possible by using historical data to train the model for characterizing design attributes and tell which individual with certain characteristics, that has been already classified, is keen to select specific features.

In the last stage, it was necessary to encompass optimization tools for obtaining accurate results. In this sense, we included GAs for selecting design features as a

complementary part of the analysis, since we moved to more complex datasets that involved a mix of categorical and numerical attributes, which in the past were not considered for applications. A collection of data that involved large quantities of variables or predictors lead us to integrate an effective way of pattern exploration, significant information, and significant interactions for customer behaviour. In the end, how significant was the information and the selection of design attributes, resulted in intelligent ways of customizing products.

The next chapter presents the case studies and shows how the proposed methodologies are applied to analysing the data for the i4 objective of customisation. The applications include classification, clustering, and feature selection in predicting potential customer needs and wants for the purpose of customizing production.

Chapter 5 Applications and Case Studies

The applications and case studies used to test the methodologies and frameworks shown in Chapter 4 are presented in this chapter. First, each dataset is introduced as part of section 5.1, in order to know the data in detail. Different datasets have been used and correspond to the different stages of the research work, which will be discussed in detail. After this, the motivation of using these datasets is presented in section 5.2, where we give a full description of the afore-mentioned stages is given. Followed by section 5.3 that corresponds to the data analysis, in which different subsections coincide with the given datasets as case studies. Finally, in section 5.4 is presented a discussion of obtained results.

5.1 Datasets for Applications

Different datasets were accessed to assess and test the proposed methodologies. At different stages of this research work we tried to give a solution to different challenges for personalizing design products considering customer needs and wants. First, it was the car evaluation dataset, and it was decided to analyse it because of the challenge that represented to classify categorical instances that represented design attributes of cars, and how good/bad customers were keen to accept those different attributes. The car evaluation dataset corresponded to an academic repository of data and these datasets can be modified for academic purposes, and it was only 7 different variables to analyse. After analysing this data, the automobile dataset was accessed from the same academic repository but since there were more variables to analyse, this represented a bigger challenge. Once finished with the datasets accessed from the academic repository, we decided to move to real data that represented a set of historical records and information to be analysed to understand customers' behaviour in specific cases. The fuel economy dataset suited perfectly to address this challenge, and it was the first time that part of the challenge was how to analyse raw data and how to determine significant information to train a

classification model that understand customer needs and wants. Following the same example as the fuel economy dataset, we decided to analyse a dataset that had the same complexity in terms of dealing with raw data and able to tell us insights from a set of historical records. The CPU dataset involves raw data and represents a collection of design attributes that can tell us what manufacturers and customers should pay attention to.

The next subsections describe in full the details about the previous-mentioned datasets.

5.1.1 Car Evaluation Dataset

This dataset was accessed from a trained data found in a machine-learning repository [124] in order to run some tests, the information of the data shows an evaluation model of cars by acceptability, overall price, buying price, price of maintenance, technical characteristics, comfort, number of doors, persons capacity to carry, and safety of the car. The dataset comprises of 1728 instances and each record contains the subsequent attributes: safety, capability describing the persons to hold, buying price, maintenance price, number of doors, the dimensions of baggage boot, and car acceptance. For this data set, the attribute of car acceptance is a category label used to classify the level of the car that customers accept, then different attributes are seen as predictive inputs. In Table 5–1 the dataset contents are presented.

Table 5–1 Car evaluation dataset [124]

Attribute name	Description	Domain
safety	Safety evaluation	Low / med / high
person	The number of passengers	2 / 4 / more
b_price	Buy Price	v-high / high / med / low
m_price	Repair price	v-high / high / med / low

size	Suitcase capacity	Small / med / big
door	The number of the door	2 / 3 / 4 / 5-more
class	level of customer acceptance	Unacc / acc / good / vgood

In order to examine the distributions for getting to know the dataset better, in Figure 5-1 the categories for the car evaluation set are presented.

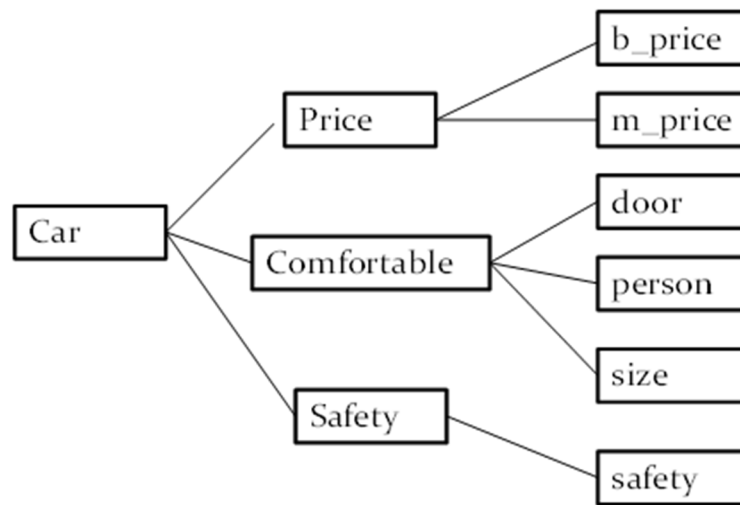


Figure 5-1 Distributions of car evaluation dataset for customization.

Following the methodology proposed in chapter 4 for the AI closed-loop, it was also considered the business problem as the following question: what reasonably cars can get good assessment? This question is first taking into account, then the obtained evaluations are used as target attributes, depending on which attribute. This is then reduced to a data mining problem, which is: find out the rules form other attributes.

5.1.2 Automobile Dataset

This dataset found in [125], consists of three types of entities: (a) the specification of an auto in terms of various characteristics, (b) is assigned insurance risk rating, (c) is normalized losses in use as compared to other cars. The second rating corresponds to the degree to which the auto is riskier than its price indicates. Cars are initially

assigned a risk factor symbol associated with its price. Then, if it is riskier (or less), this symbol is adjusted by moving it up (or down) the scale. Actuaries call this process "symboling". A value of +3 indicates that the auto is risky, -3 that it is probably safer.

The third factor is the relative average loss payment per insured vehicle year. This value is normalized for all autos within a particular size classification (two-door small, station wagons, sports/speciality, etc...), and represents the average loss per car per year. In Table 5–2 the contents of the automobile dataset are shown.

Table 5–2 Automobile data

Attribute	Attribute Range	Attribute	Attribute Range
symboling	-3, -2, -1, 0, 1, 2, 3.	curb-weight:	Continuous from 1488 to 4066.
normalized-losses:	Continuous from 65 to 256.	engine-type:	dohc, dohcv, l, ohc, ohcf, ohcv, rotor.
make	alfa-romeo, Audi, bmw, Chevrolet, dodge, Honda, Isuzu, jaguar, Mazda, Mercedes-Benz, mercury, Mitsubishi, Nissan, Peugeot, Plymouth, Porsche, Renault, Saab, Subaru, Toyota, Volkswagen, Volvo	num-of-cylinders:	Eight, five, four, six, three, twelve, two.
fuel-type	Diesel, gas.	engine-size:	Continuous from 61 to 326.
Aspiration	Std, turbo.	fuel-system:	1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
num-of-doors	Four, two.	bore:	Continuous from 2.54 to 3.94.
body-style	Hardtop, wagon, sedan, hatchback, convertible.	stroke:	Continuous from 2.07 to 4.17.
drive-wheels	4wd, fwd, rwd.	compression-ratio:	Continuous from 7 to 23.

engine-location	Front, rear.	horsepower:	Continuous from 48 to 288.
wheel-base	Continuous from 86.6 to 120.9.	peak-rpm:	Continuous from 4150 to 6600.
Length	Continuous from 141.1 to 208.1.	city-mpg:	Continuous from 13 to 49.
Width	Continuous from 60.3 to 72.3.	highway-mpg:	Continuous from 16 to 54.
height	Continuous from 47.8 to 59.8.	price:	Continuous from 5118 to 45400.

This dataset comprises 205 instances and 26 attributes, as shown in Table 5–2.

5.1.3 Fuel Economy Dataset

The dataset was accessed from the fueleconomy.gov website [126], run by the U.S. Department of Energy’s Office of Energy Efficiency and Renewable Energy. The U.S. Environmental Protection Agency lists different estimates of fuel economy for passenger cars and trucks. For each vehicle, various characteristics such as engine displacement or number of cylinders were recorded. Along with these values, laboratory measurements were taken for the city and highway miles per gallon (MPG) of each vehicle.

The accessed fuel economy dataset includes a collection of characteristics and measures made by [126], from the year 2014 to 2016. They are used to create a prediction for 2017. In Table 5–3, each attribute description and attribute type are presented.

In [126], information for many past years could be accessed, but only information from 2014 to 2016 was used to train the data using machine learning approaches and get a prediction from the training set. Figure 5-2 presents a brief comparison of each unit considered for the average of fuel spent, the average CO₂ emissions inside cities across U. S states, and an average of annual fuel cost on conventional fuel.

It is suggested in [127] that when trying to build a predictive model it is best to identify single predictors. Hence, it was selected as predicting variables all the 52 different attributes, in which the variables “spend of fuel over the last five years” and “city CO₂ rounded adjusted emission” were outlined as significant predictors. The selected response was the manufacturer’s name or brand.

Table 5—3 Description of fuel economy data attribute.

Attribute	Attribute Type	Attribute	Attribute Type	Attribute	Attribute Type
Model Year	Numerical	# Cyl	Numerical integer	Carline	Mix of numerical and categorical
Mfr Name	Categorical	Division	Categorical	Eng Displ	Numerical
Transmission	Categorical	City FE (Guide) - Conventional Fuel	Numerical integer	Hwy FE (Guide) - Conventional Fuel	Numerical integer
Comb FE (Guide) - Conventional Fuel	Numerical integer	City Unadj FE - Conventional Fuel	Numerical	Hwy Unadj FE - Conventional Fuel	Numerical
Comb Unadj FE - Conventional Fuel	Numerical	City UnrdAdj FE - Conventional Fuel	Numerical	Hwy UnrdAdj FE - Conventional Fuel	Numerical
Comb UnrdAdj FE - Conventional Fuel	Numerical	Guzzler?	Categorical	Air Aspir Method	Categorical
Trans	Categorical	Trans, Other	Categorical	#Gears	Numerical integers
Lockup Torque Converter	Y, N	Trans Creeper Gear	Y, N	Drive Sys	Categorical

Max Ethanol % - Gasoline	Numerical integers	Fuel Usage - Conventional Fuel	Categorical	Fuel Unit - Conventional Fuel	Categorical
Gas Guzzler Exempt (Where Truck = 1975 NHTSA truck definition)	Categorical	Annual Fuel Cost - Conventional Fuel	Numerical integers	EPA Calculated Annual Fuel Cost - Conventional Fuel --- Annual fuel cost error.	Mix of categorical and numerical
Intake Valves Per Cyl	1, 2	Exhaust Valves Per Cyl	1, 2	Carline Class	Categorical
Car/Truck Category - Cash for Clunkers Bill.	Categorical	Calc Approach Desc	Categorical	Release Date	Numerical integers
EPA FE Label Dataset ID	Numerical integers	Fuel Metering Sys Cd	Categorical	\$ spent over five years (increase in fuel costs over five years - on label)	Numerical integers
City CO ₂ Rounded Adjusted	Numerical integers	Hwy CO ₂ Rounded Adjusted	Numerical integers	Comb CO ₂ Rounded Adjusted (as shown on FE Label)	Numerical integers
Oil Viscosity	Categorical				

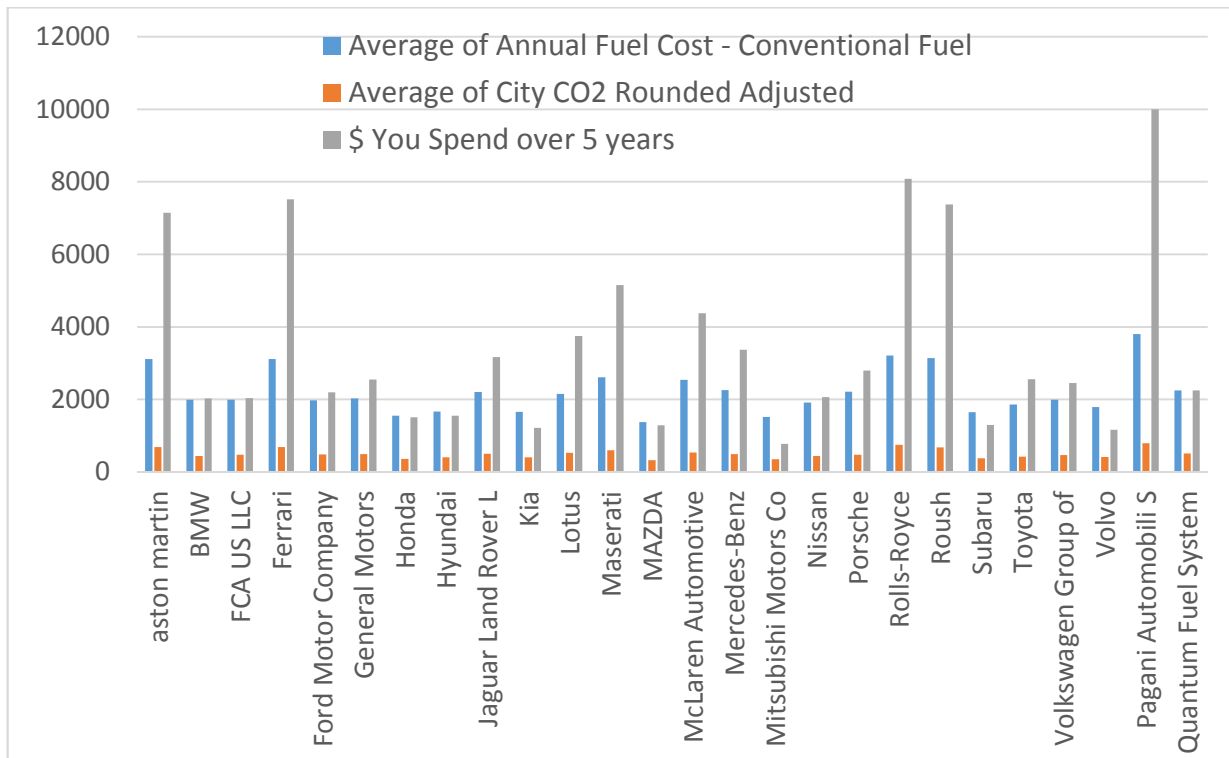


Figure 5-2 Relation of fuel economy dataset for the average USD spent, the average CO₂ emissions, and the average annual fuel costs of conventional fuel.

5.1.4 CPU Dataset

It was considered an application that contained detail specifications, costs, release dates, amongst other characteristics of computers and components. Because of the nature of this dataset, it matches perfectly with the description of the problem we wanted to address in this work. This dataset comprises a collection of data based on Central Processing Units (CPUs) components, published by [128]. In this collection of data there are 45 different columns or variables that involve: product collection, vertical segment, processor number, lithography, recommended customer price, number of cores, number of threads, processor base frequency, max turbo frequency, cache, bus speed, thermal design power, embedded options available, conflict free, max memory size, memory type, max number of memory channels, max memory bandwidth, error-correcting code (ECC) memory supported, processor graphics, graphics base frequency, graphics max dynamic frequency, graphics video max

memory, graphics output, support 4k, max resolution HDMI, max resolution display port (DP), max resolution embedded display port (eDP) integrated flat panel, direct X support, peripheral component interconnect (PCI) express, PCI express configurations, max number of PCI express lanes, temperature, intel hyper threading technology, Intel Virtualization Technology VTx, intel 64, instruction set, instruction set extensions, idle states, thermal monitoring technologies, secure key, and execute disable bit. In total this dataset contains 2283 rows for each column, and Table 5—4 gives in detail the content of each attribute.

These attributes were used to predict what customers are most likely to consider for a CPU design based on historical data. This helps manufacturers like Intel among others to decide the best design and what direction new products should take, but as well as customers to select the best choice based on their needs and wants. In this data are included computer components that involve specifications that manufacturers consider as design elements when creating (manufacturing) products of this kind. A wide range of components considered by manufacturers when building a CPU is reviewed by an individual that wants to place a purchase, and the aim of this analysis is to be able to recommend a set of design features that suit best for each individual and make predictions based on previous decisions to constantly improve the recommendation system. It was selected as a response the variable “product collection”, all the other variables were used as predictors.

In Figure 5-3 the relation of customer price attribute and the product collection (model) are depicted. This plot shows how expensive the processor models can be compared to each other.

Table 5—4 Description of CPU data attribute.

Attribute	Attribute Type	Attribute	Attribute Type	Attribute	Attribute Type
Product Collection	Categorical	Conflict Free	Y, N	DirectX Support	Numerical

Vertical Segment	Categorical	Max Memory Size	Categorical	OpenGL Support	Categorical
Processor Number	Categorical	Memory Types	Categorical	PCI Express Revision	Numerical integer
Status	Categorical	Max nb of Memory Channels	Numerical	PCI Express Configurations	Categorical
Launch Date	Categorical	Max Memory Bandwidth	Categorical	Max nb of PCI Express Lanes	Numerical
Lithography	Numerical	ECC Memory Supported	Y, N	Temperature	Categorical
Recommended Customer Price	Numerical	Processor Graphics	Categorical	Intel Hyper Threading Technology	Y, N
Nb of Cores	Numerical	Graphics Base Frequency	Categorical	Intel Virtualization Technology VTx	Y, N
Nb of Threads	Numerical	Graphics Max Dynamic Frequency	Categorical	Intel 64	Y, N
Processor Base Frequency	Categorical	Graphics Video Max Memory	Categorical	Instruction Set	Categorical
Max Turbo Frequency	Categorical	Graphics Output	Categorical	Instruction Set Extensions	Categorical
Cache	Categorical	Support 4k	Categorical	Idle States	Y, N
Bus Speed	Categorical	Max Resolution HDMI	Categorical	Thermal Monitoring Technologies	Y, N

Thermal Design Power	Numerical	Max Resolution DP	Categorical	Secure Key	Y, N
Embedded Options Available	Y, N	Max Resolution eDP Integrated Flat Panel	Categorical	Execute Disable Bit	Y, N

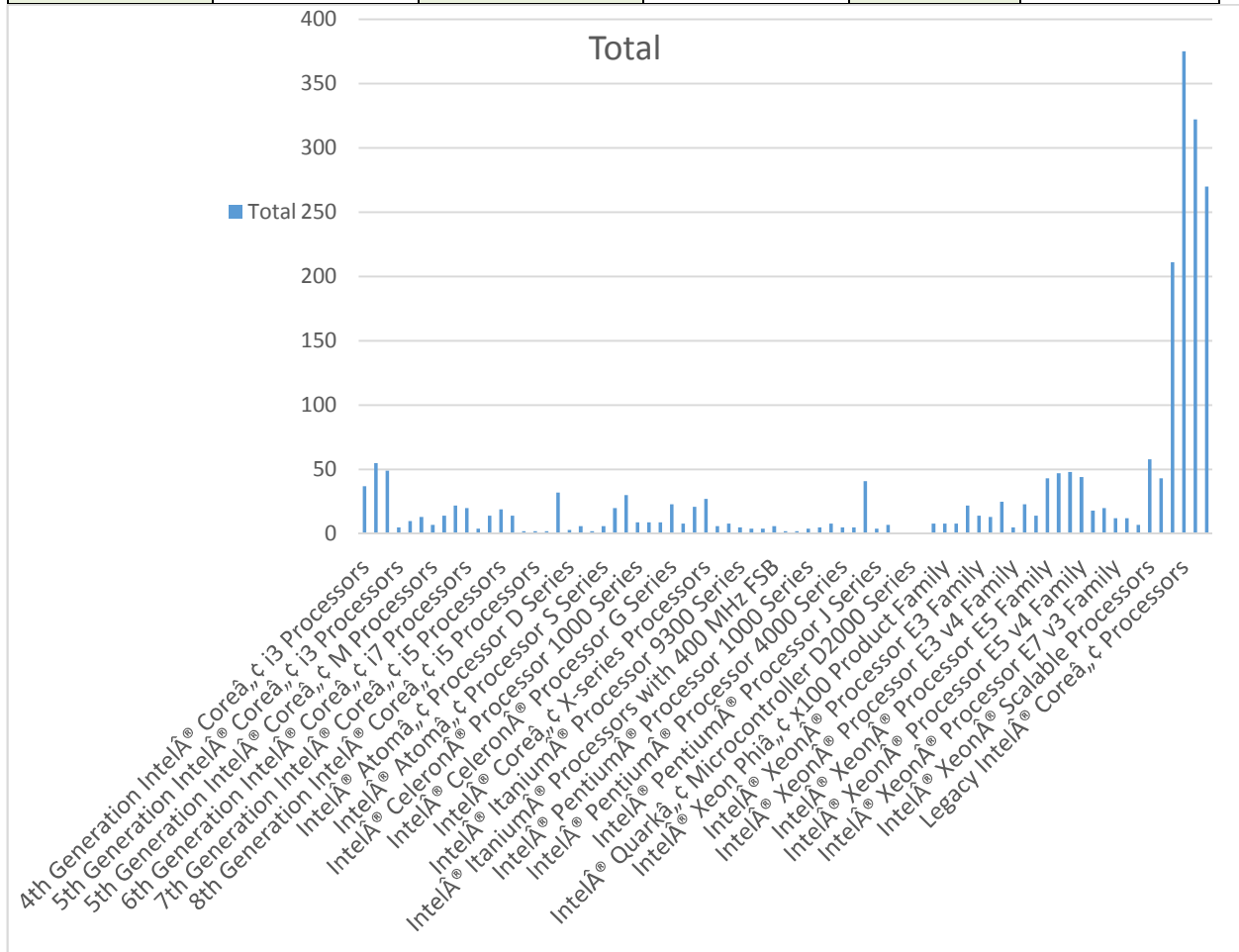


Figure 5-3 Relation of recommended customer price attribute vs product collection or models attribute.

5.2 Selected Case Studies to Illustrate the Applications

As stated before, once completed the process of proposing a closed-loop framework, we decided to test the methodology using case studies for validating that the proposed

framework actually was giving a desirable solution for predicting customer needs and wants.

In the case of car evaluation and automobile datasets, the data represented a very simple challenge since the data was obtained from academic repositories, i.e. data that was already manicured for academic purposes and a special treatment for analysing it was given in terms of some of the fields were already pre-processed and the data to some extent was already trained successfully. Still, the attributes match with the results we were trying to obtain from the data analysis. The datasets involved instances either categorical or numerical that represented design attributes, so the assignation of classes/categories, pattern recognition, and selection of features was tested to see how was in practice testing the proposed approach.

Moving forward with the validation of the proposed framework, we decided to use the fuel economy dataset since it represented a bigger challenge. The analysis of this dataset helped us to refine the framework and include a more complete analysis. Since the source of the set involved raw data that needed to be processed, this was also considered as part of the methodology.

The CPU dataset involved specific design attributes, specifications, and various characteristics for computers that manufacturers use when designing a computer. Thus, being able to predict which specific characteristics individuals might choose from the whole range of computer components was the motivation to use this application. It was discovered in practice that part of the challenge when dealing with historical data of this nature also involves pre-processing since this was raw data that required a certain level of arrangements before analysing it.

5.3 Data Analysis

In this section are comprised the data analysis results obtained when performed the classification, cluster, and feature selection analysis to the different case studies. The order presented corresponds to the chronological order of accessing the data, analyse it, and publish results in papers. As discussed before, each analysis helped us to improve the proposed frameworks, from an early stage in which it was clear that customization needed to be obtained from a predictive closed-loop, then moving forward we discovered that machine learning can actually deal with design attributes if the right analysis is conducted. With the use of machine learning approaches to training models for predicting customer needs and wants, it was discovered that a specific method can lead to an incomplete analysis and for this is better to consider a combination of approaches.

5.3.1 Car Evaluation Dataset Results

Self-Organizing Map

This dataset was first analysed using SOM as part of the unsupervised learning or cluster analysis, as discussed in chapter 3, section 3.1. The following tables and figures represent the analysis obtained when using SOM approach. In Figure 5-4 the results of the SOM cluster analysis are shown, in which all the weights connect to each other, then compete (pattern recognition), and finally cooperate to create the neighbourhood, as part of the process presented in chapter 3, section 3.1. Figure 5-4 provides evidence about neighbourhoods created: the darker colours represent larger distances, and the lighter colours represent smaller distances, these neighbours give us inside about the adaption process in which the self-organizing feature creates the map displayed [129]. For which the first neuron in the inferior corner on the left results to be the strongest one, meaning that attribute selected is “safety”, if the input “safety” is low it will directly fall under unacceptable (“unacc”). Whatever

estimation of safety is, if “person” value is 1, the entry will fall under unacceptable. This is represented in the right part of the figure presented below, where it shows the assigned clusters.

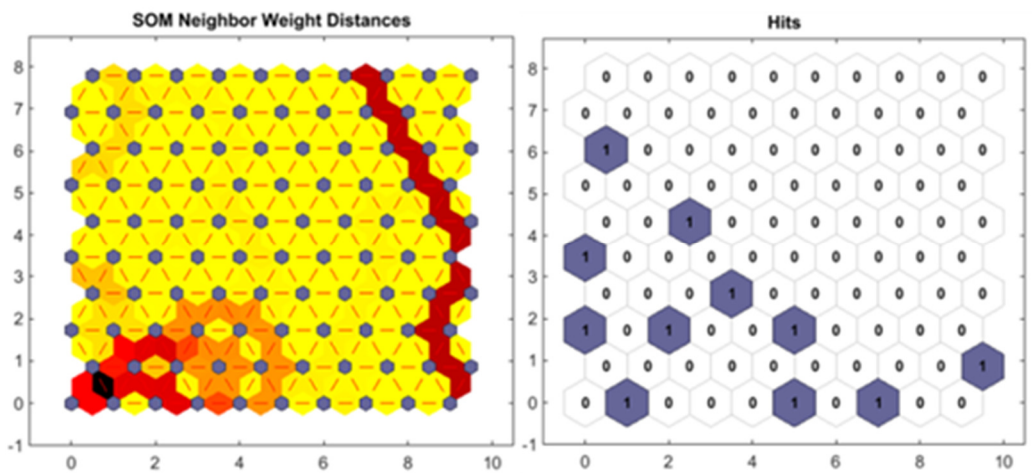


Figure 5-4 Results of tested data. SOM weight distances on the left, and SOM clusters found on the right.

Here, the SOM work with 10 hidden neurons, and 200 iterations. The confusion matrix is shown in Table 5–5 resulted from the analysis with Matlab.

Table 5–5 Confusion matrix for the SOM

a	b	c	d	Classified
1171	28	0	3	a= unacc
7	292	4	9	b= acc
0	0	44	0	c= vgood
0	5	5	37	d= good

Then, it was also tested the average clustering coefficient with a value of 0.833. This means the degree to which nodes in a graph tend to cluster together. Meaning that from the clusters 4/5 can be clustered together. To test the accuracy of the model obtained, Table 5–6 shows in detail each class evaluated, this table as well is created from the Matlab analysis.

Table 5–6 Model accuracy by class

TP Rate	FP Rate	Precision	Recall	F-Measure	Classified
0.974	0.017	0.994	0.974	0.984	unacc
0.936	0.026	0.898	0.936	0.917	acc
1	0.006	0.83	1	0.907	vgood
0.787	0.008	0.755	0.787	0.771	good

From the results presented above, it can be inferred that the model performs good, from all assessing values followed with less serious miss-classification, that there were sixty-one entries that show wrong classification, it can be told from Table 5–5 that even those values are in a wrong category, most of them are leading a category close to their actual categories. Part of the weight adaption when training this dataset can be concluded that the set neurones that best self-adapted were the “vgood”, as presented in Table 5–6, the approximation shows that the rate obtained of 1 provides evidence of featured map became member of this neighbourhood. This analysis represents part of the closed-loop proposed in Chapter 4, section 3, Figure 4-3; where dataset was accessed, and this step is identified in block number 2 of the aforementioned figure as the exploration and data analysis process.

Cluster k-means

Clustering was obtained using the WEKA toolbox in Matlab, in this case, it was tested against the simple k-means scheme for better results. Table 5–7 shows the results when evaluating the model using simple K-means. Two clusters were selected and 10 seeds and the training set was used to run the algorithm, where all 7 attributes and 1728 instances were considered.

Table 5–7 Simple k-means clustering testing 7 attributes.

Instances	Percentage	Classified
1104	64%	Cluster 0
624	36%	Cluster 1

Following this, classification of clusters, as the selected mining method to build the model was also performed. Classifier decision tree (ID3) was selected because it uses greedy strategy to select the best attribute by splitting the dataset on each iteration as discussed in [130]. This algorithm helps us to select the best attribute, thanks to gain information displayed on each generated node. The results of the model accuracy by class attribute are presented in Table 5–8.

Table 5–8 Model accuracy by class

TP Rate	FP Rate	Precision	Recall	F-Measure	Classified
1	0	1	1	1	unacc
1	0	1	1	1	acc
1	0	1	1	1	good
1	0	1	1	1	vgood

In TABLE 5–9 the confusion matrix for this classification using the ID3 method in Weka toolbox is presented.

TABLE 5–9 Confusion matrix for classified attributes.

a	b	c	d	Classified
1210	0	0	0	a= unacc
0	384	0	0	b= acc
0	0	69	0	c= good
0	0	0	65	d= vgood

For this classification, there were no incorrect classified instances, and when run in Matlab the time to build the model was 0.01 seconds. It can be inferred from Table 5–8 and TABLE 5–9 that the classification model works well - instances assigned to domain unacceptable (unacc) turns to be the ones that have more impact on cluster assignation with a value of 1210 instances. Simple k-means clustering and ID3 show that classification of domains for each attribute can reflect the exact quantity of clusters. On the other hand, attributes like maintenance, buying and doors show more

incorrect clustered instances, as presented in Figure 5-5 and Figure 5-6, where the interaction of these variables is shown in the scatter plot, and the “x” mark represents the incorrect instances classification.

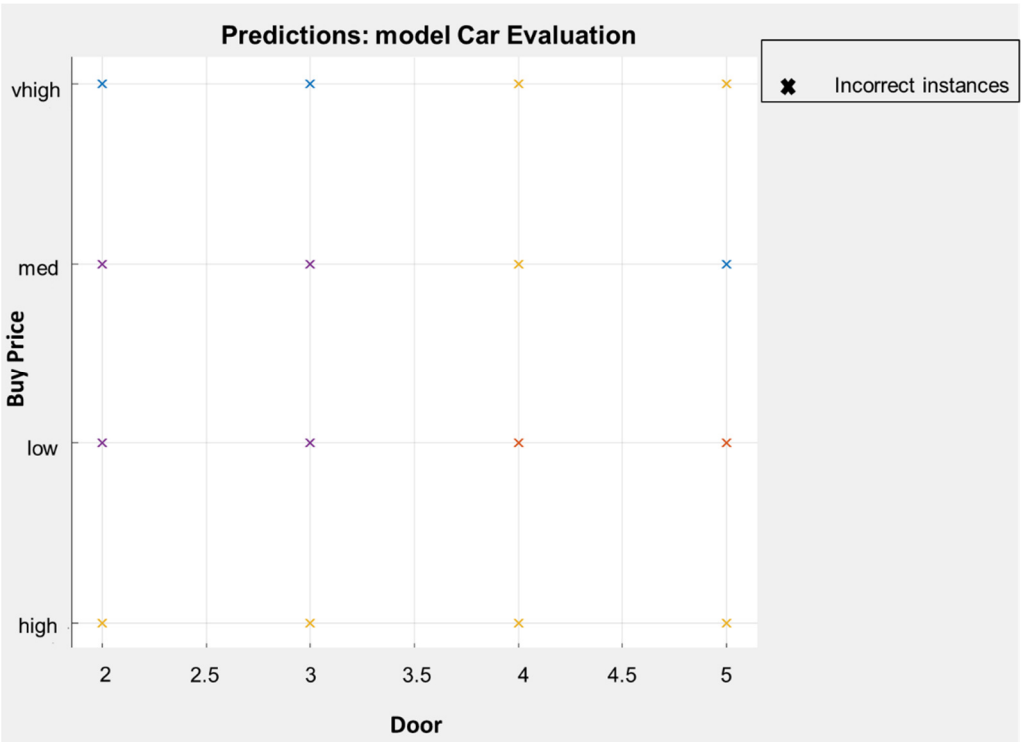


Figure 5-5 Scatter plot for the incorrect classified instances of variables “buy price vs “doors” using simple k-means.

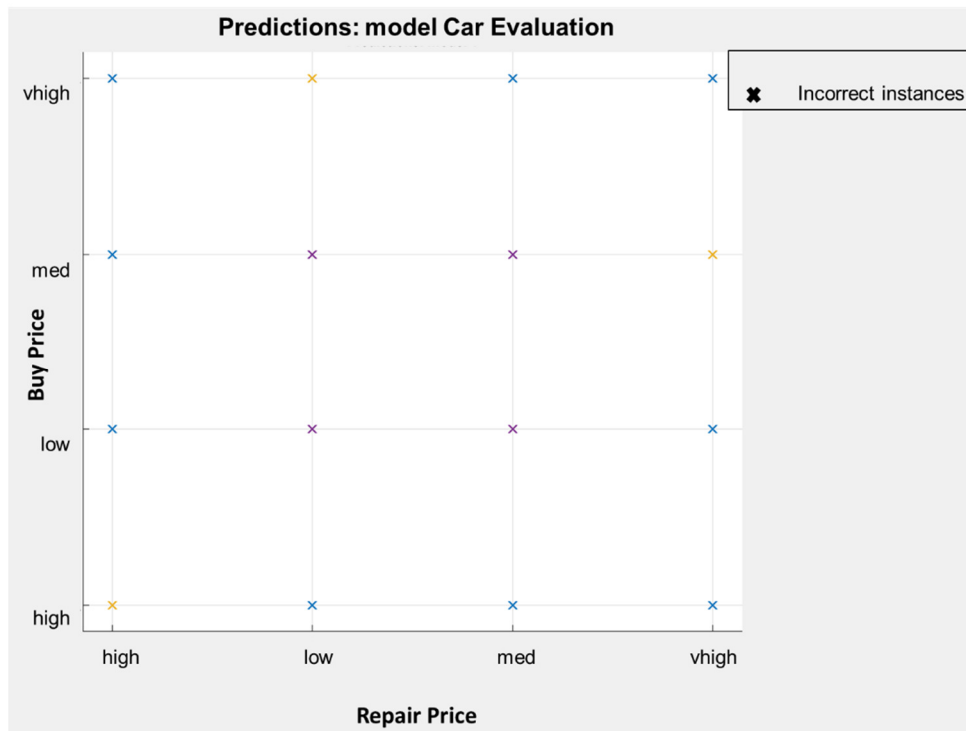


Figure 5-6 Scatter plot for the incorrect classified instances of variables “buy price vs “repair price (maintenance)” using simple k-means.

According to [131] *k*-means clustering can represent weaknesses: a) With fewer samples of data, initial grouping will determine the cluster significantly; b) The number of clusters, *k*, must be determined beforehand; c) With fewer samples of data, inaccurate clustering can occur; d) It cannot be inferred which variable contributes more to the clustering process since it is assumed that each has the same weight; e) The accuracy of mathematical averaging weakens because of outliers, which may pull the centroid away from its true position; and f) The results are clusters with circular or spherical shapes because of the use of distance.

Feature Selection Using Genetic Search

Once clusters were found, the following step was to use Coefficients Subset Evaluation (CfsSubsetEval) that according to [131] and libraries inside WEKA toolbox, means that: evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. And as a

search method, it was used Genetic Search, with a probability of search equals to 0.6, a maximum of generations of 20, mutation probability of 0.033, population size of 20, report frequency of 20, number of seeds equals to 1 and starting set number 1. In [132] it says that for every single application or experiment, search algorithms can have different settings when dealing with test data generation, and therefore needs to be empirically tested to find the right combination of settings that work for your tested dataset. In this sense, the parameters used for the genetic search were determined on the performance and with the objective of gaining time on algorithm runtime, which it was decided to sacrifice number of generations and population size and use a lower number, but testing it against greater numbers, and the results showed no difference between greater number of generations and population size. For the mentioned criteria, it was disregarded the attribute class. The results obtained show that safety was the best attribute. Different to what it was obtained for class attribute, disregarding safety as the main attribute, it was selected with a higher level of prediction the class attribute. Disregarding all the remaining attributes (buying, paint, doors, persons, and lug_boot) present the same selection: class. Based on the results presented above, it is clear that the best-selected attribute is class.

5.3.2 Automobile Dataset Results

Fuzzy c-means Clustering

The results of the fuzzy c-means are shown in Figure 5-7. Here, the partition of the 3 clusters can be noticed. The scatter plot shows the connections between all the instances. From here, Matlab function for fuzzy c-means update the cluster centres and membership grades of each data point, clusters are iteratively moved from the centre to the right location inside the dataset. The selected parameters for the fuzzy c-means were 3 clusters, exponent =3, the maximum of iterations = 100, and minimum improvement= 1e-05. Since iterations are based on minimizing an objective function that represents the distance from any given data point to a cluster centre weighted

by that data point's membership grade. Membership function plots obtained are presented in Figure 5-8, here for each cluster shows when it reached the maximum of iterations, or when the objective function improvement between two consecutive iterations is less than the minimum amount of improvement specified. For the given dataset, the considered attribute to build the membership functions was “price” variable. The values found in “price” range from \$5’118 to \$45’400, and were classified into five fuzzy sets (very low, low, medium, high, and very high), where 3 clusters were found as shown in Figure 5-7.

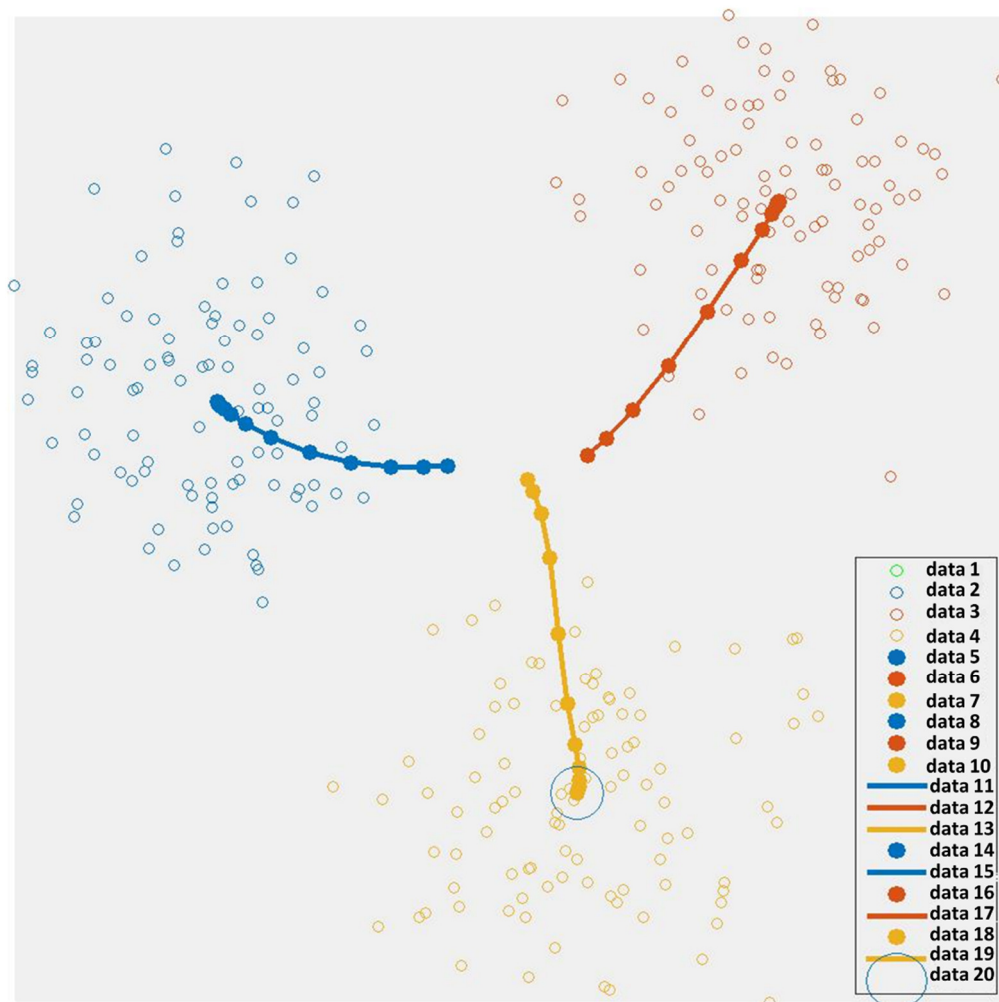
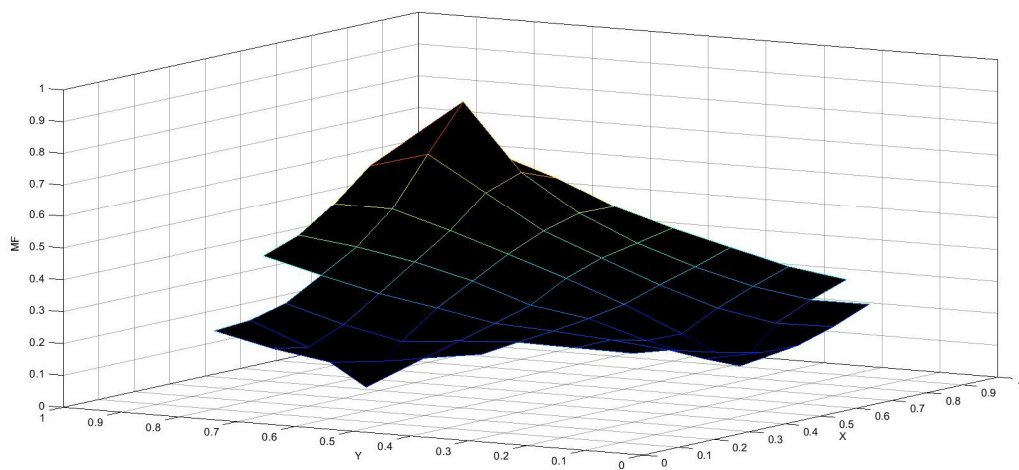
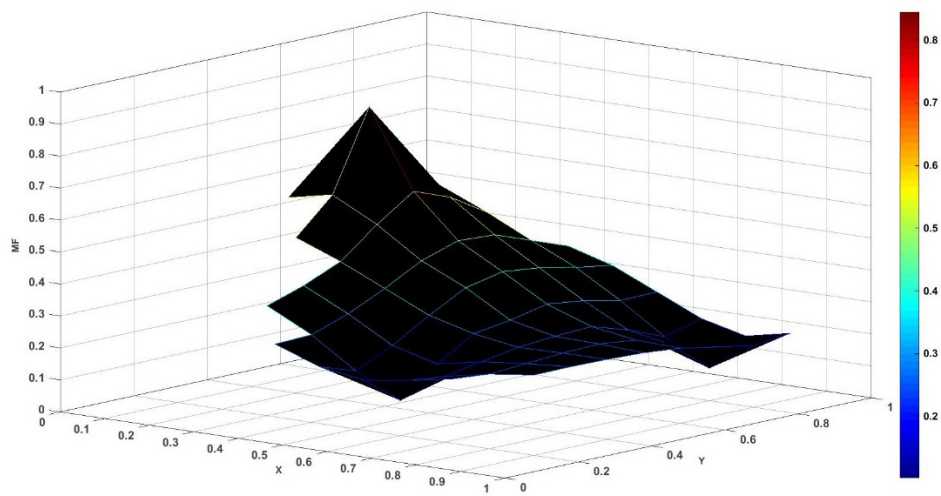


Figure 5-7 Results of tested data. Fuzzy c-means with 3 clusters found.



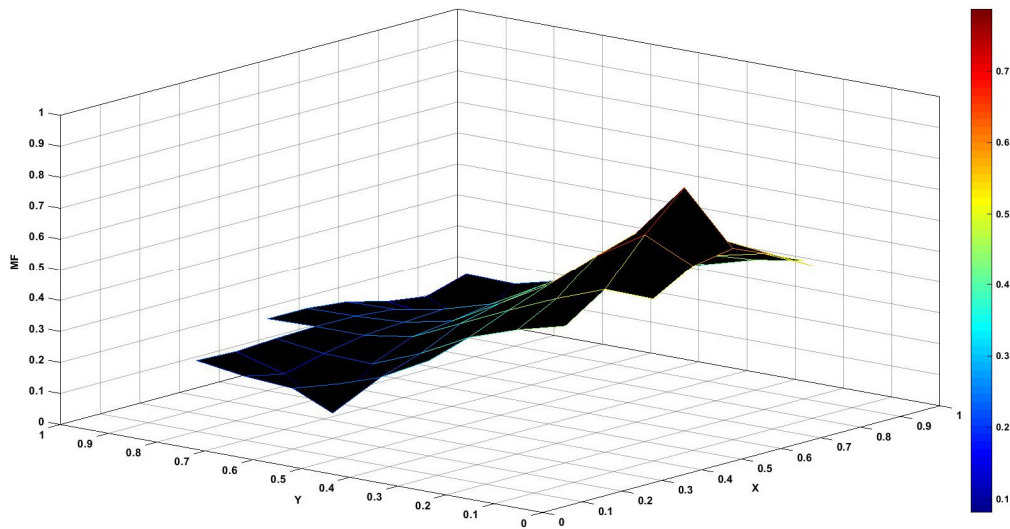


Figure 5-8 Membership function. From top to bottom: cluster 1, 2 and 3 results.

Attribute Classification

Once the clustering was done, it was processed the training data to obtain the attribute classification inside Matlab toolbox for machine learning, where it was as well embedded parallel routine for speeding up the whole process. Testing with several classifier algorithms, the results are presented in Figure 5-9.

The confusion matrix presented in Figure 5-9 helps to assess the classifier performance, in which this plot was used to understand how the currently selected classifiers obtained the desired performance in each class. The confusion matrix helps to identify the areas where classification was performed poorly. All those values coloured in green show the corrected classified instances, based on the attribute that best reflected the desired selection: manufacturer or make. The red slots represent the incorrect instances. Here the manufacturer (make) was selected as the predictive variable in order to provide which of the observed brands are more attractive to customers based on all the considered variables.

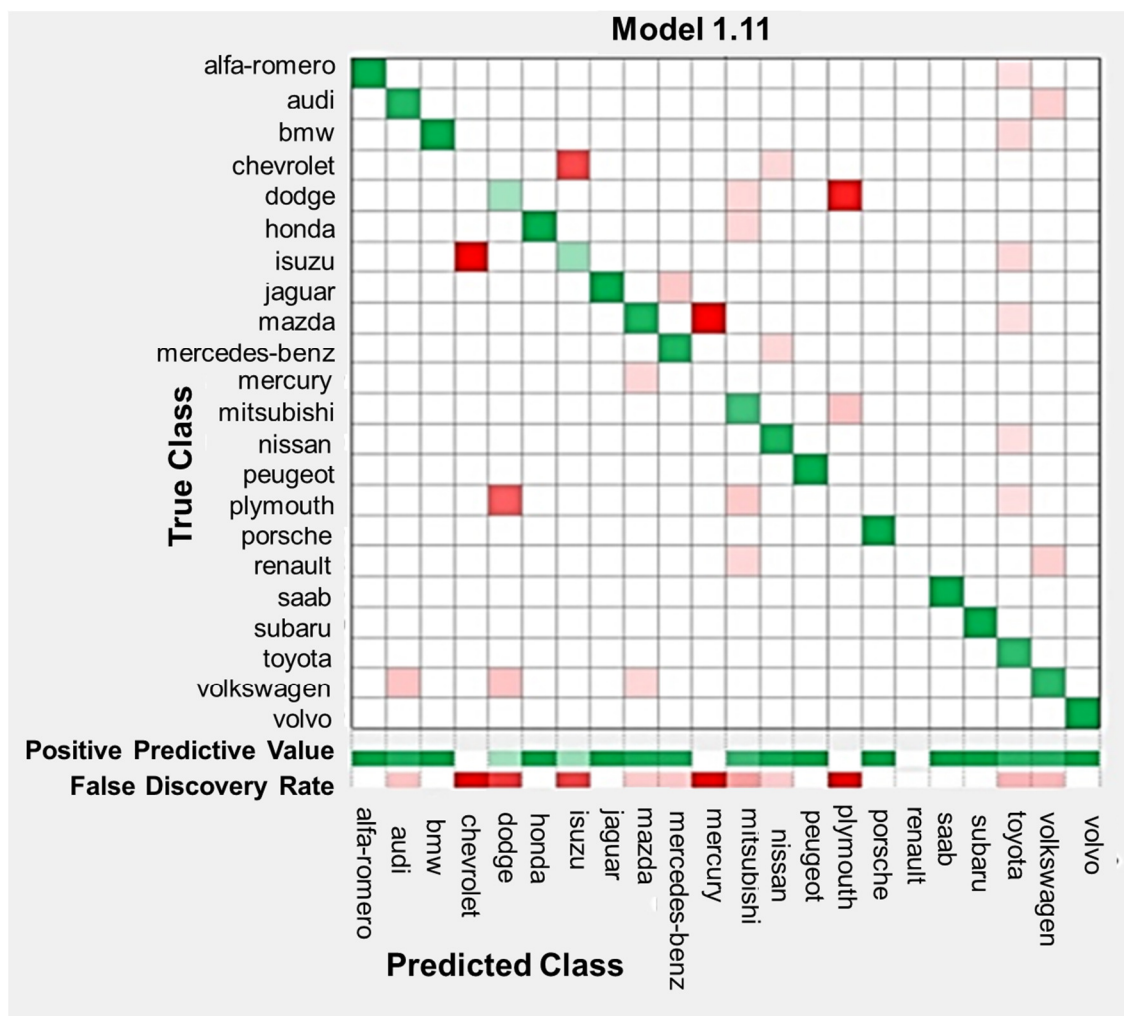


Figure 5-9 Confusion matrix obtained for positive predictive values.

For the presented plot in Figure 5-10 can be inferred what type of attributes represent the most corrected classified instances to the predictive model. The parallel coordinates plot helps to understand relationships between features and useful predictors for separating classes, where the standardized values are used to see the distribution of the predictors (make) along the mean distribution on the interaction between each feature. The selected response variable was the Manufacturer, and each colour represents the brand related to the predictors (fuel-type, number of doors, body style, engine locations, HP, etc.). For which the strongest relation is found with the engine location, number of cylinders and the HP variables. Moreover,

once the attribute selection was performed using the GA selection, it was selected the following instances: num-of-doors, drive-wheels, height, engine-type, num-of-cylinders. Those were performed with a crossover probability of 0.6, a max of generations of 20, mutation probability of 0.033, initial population size of 20, and an initial seed. The parameters were determined, as stated earlier in this chapter, by empirically testing the initial settings and obtain the minimum runtime possible for the search algorithm.

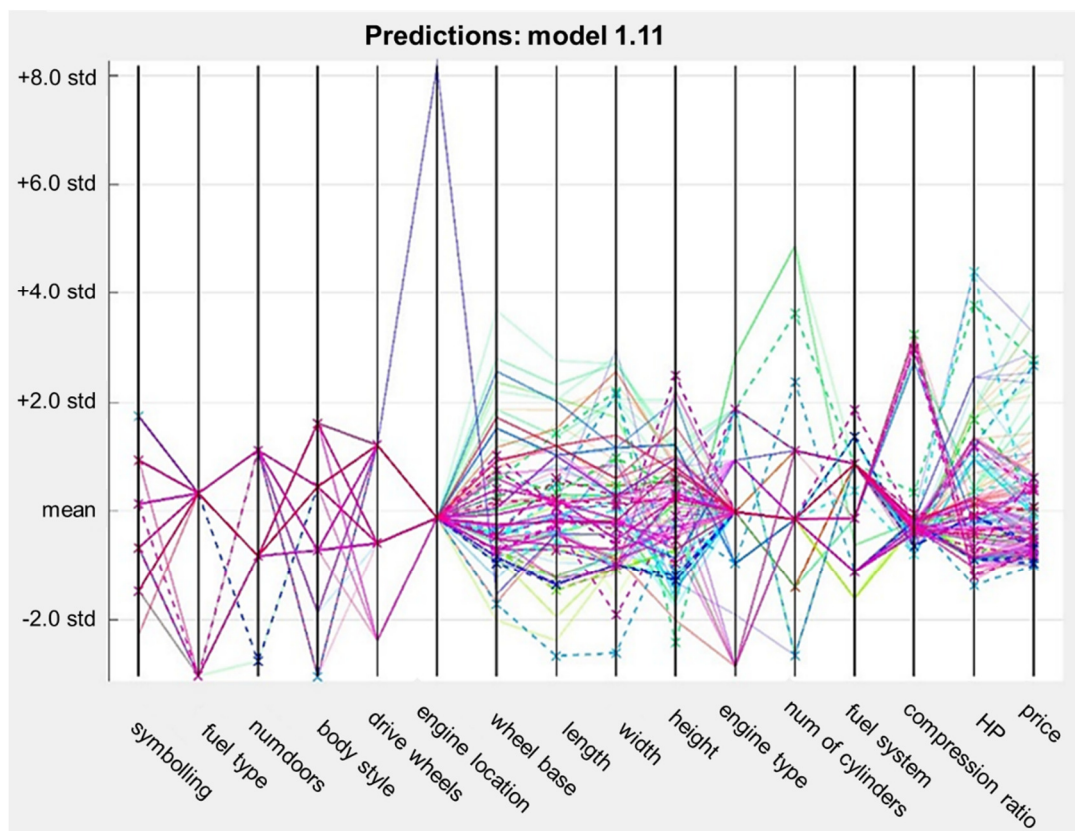


Figure 5-10 Parallel coordinates plot for membership functions.

5.3.3 Fuel Economy Dataset Results

In this section, the results obtained from the fuel economy dataset are presented. We imported the fuel economy dataset using the Matlab classification learner app. It was selected as main predictors the variables “spend of fuel over the last five years”,

“city CO₂ rounded adjusted emission”, and “manufacturer name” as a response. One of the reasons why this application was selected is because of the automation feature that enables us to run several parallel classifiers and see which can obtain the best predictive model. The dataset encompassed 52 attributes split into 23 categorical and 29 numerical ones. The total instances considered for this dataset was 4655.

Attribute Classification

In Figure 5-11 to Figure 5-17, the scatter plots for each classification approach tested for the fuel dataset are depicted, where the correct and incorrect instances obtained from each classification approach are shown, as well as interactions between variables and interactions. Figure 5-11 presents the corresponding instances classified correctly for the decision tree classifier. In this figure, we only considered vehicles made by the following companies: BMW, Chrysler Group LLC, FCA US LLC, General Motors, Mazda, Mercedes-Benz, Nissan, Rolls-Royce, Toyota, and Volvo. In Figure 5-12, similar to the previous figure, the correct classified items are depicted.

The instances considered were: Ford Motor Company, Maserati, Mitsubishi Motors Co, Porsche, Subaru, and Volkswagen Group. For better visualization purposes we decided to break down the same classification using different manufacturing names (instances). The instances classified correctly presented in Figure 5-13 are: Ferrari, Honda, McLaren Automotive, Pagani Automobili S, Quantum Fuel System, Roush, Subaru, Volkswagen, and Aston Martin. The incorrectly classified instances are depicted in Figure 5-14: Volkswagen Group, Volkswagen, and Aston Martin. The corresponding manufacturer colour was identified as well, and as mentioned previously the reason why it was decided to show different plots for different instances is for a better visualization.

A comparison of variables spent of last five years of fuel vs the annual cost of conventional fuel is presented in Figure 5-15 and considers the following instances:

BMW, Chrysler Group LLC, FCA US LLC, General Motors, Mazda, Mercedes-Benz, Nissan, and Toyota. In Figure 5-16 presents the other correctly classified instances: Ford Motor, KIA, Maserati, Mitsubishi Motors Co, Subaru, and Volkswagen Group. In both plots (Figure 5-15 and Figure 5-16), decision trees were used as well, and all the correctly classified instances were identified. Figure 5-17 shows the incorrect instances for the variable Spent over the five years vs annual fuel cost. The considered instances that presented incorrect classification were as follows: Audi, General Motors, Maserati, Volkswagen Group, Aston Martin. Finally, Figure 5-18 and Figure 5-19 show a comparison of the variables, spent of last five years of fuel vs the use of fuel in the city. These plots were obtained using the SVM classifier. In both plots, only 1 colour was identified, which corresponds to General Motors. Considering the entire selection of manufacturers' names, including the correct and incorrect instances, only displaying the General Motors manufacturer makes this result non-desired.

For the scatter plots, the following range of colours was used to identify each manufacturing name contained in the fuel economy dataset:

■ Audi, ■ BMW, ■ Bentley, ■ Bugatti, ■ Chrysler Group LLC, ■ FCA Italy, ■ FCA USA LLC, ■ Ferrari, ■ Ford Motor Company, ■ General Motors, ■ Honda, ■ Hyundai, ■ Jaguar Land Rover, ■ Kia, ■ Lamborghini, ■ Lotus, ■ Mazda, ■ Maserati, ■ McLaren Automotive, ■ Mercedes Benz, ■ Mitsubishi Motors, ■ Mobility Ventures, ■ Nissan, ■ Pagani Automobili, ■ Porsche, ■ Quantum Fuel System, ■ Rolls-Royce, ■ Roush, ■ Subaru, ■ Toyota, ■ Volkswagen, ■ Volvo, and ■ Aston Martin.

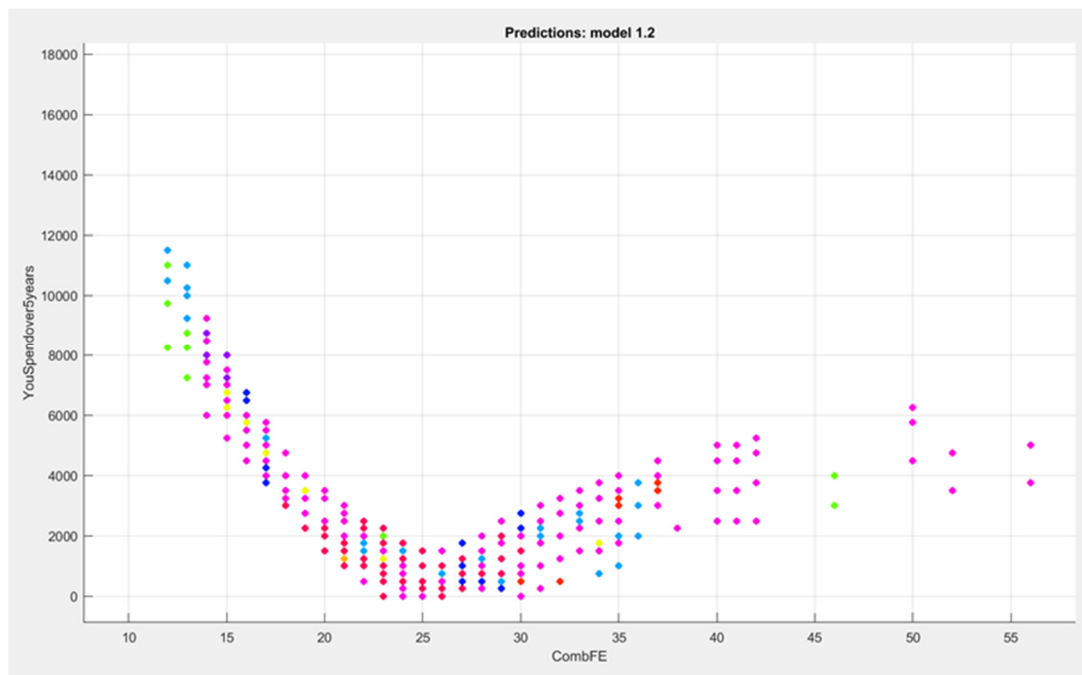


Figure 5-11 Scatter plot for the correct instances using the decision trees classifiers of variable “spent of last five years of fuel” (measured in \$USD) vs “use of fuel in the city” (measured in miles per gallon). Considered instances: BMW, Chrysler Group LLC, FCA US LLC, General Motors, Mazda, Mercedes-Benz, Nissan, Rolls-Royce, Toyota, and Volvo.

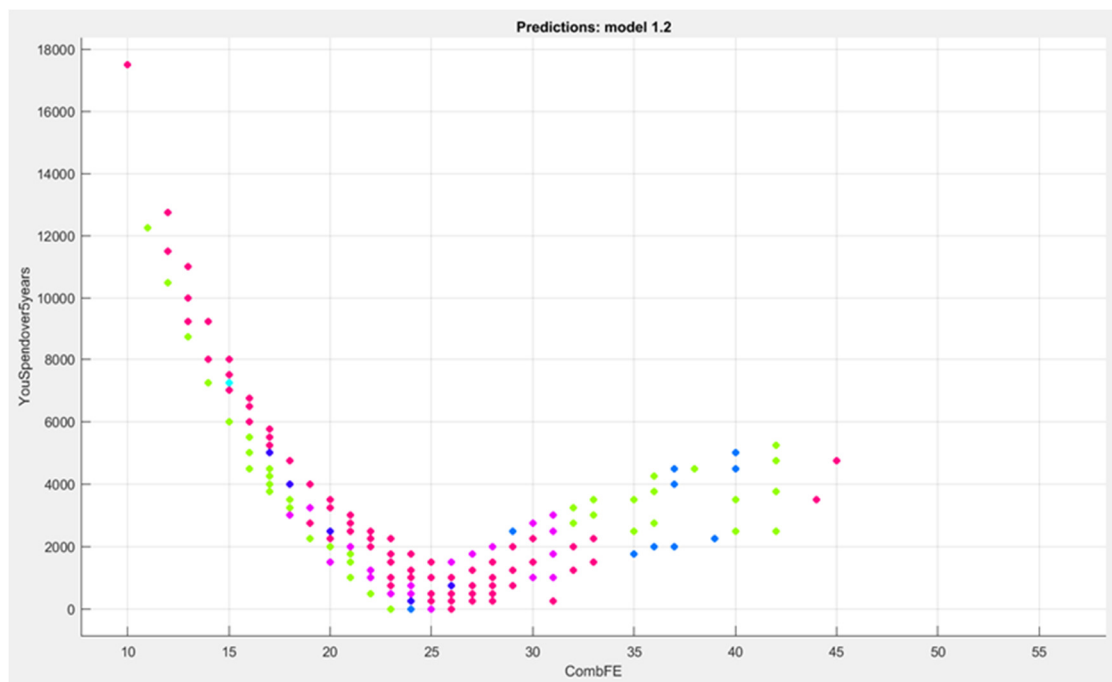


Figure 5-12 Scatter plot for the correct instances using the decision trees classifiers of variable “spent of last five years of fuel” (measured in \$USD) vs “use of fuel in the city” (measured in miles per gallon). Considered instances: Ford Motor Company, Maserati, Mitsubishi Motors Co, Porsche, Subaru, and Volkswagen Group.

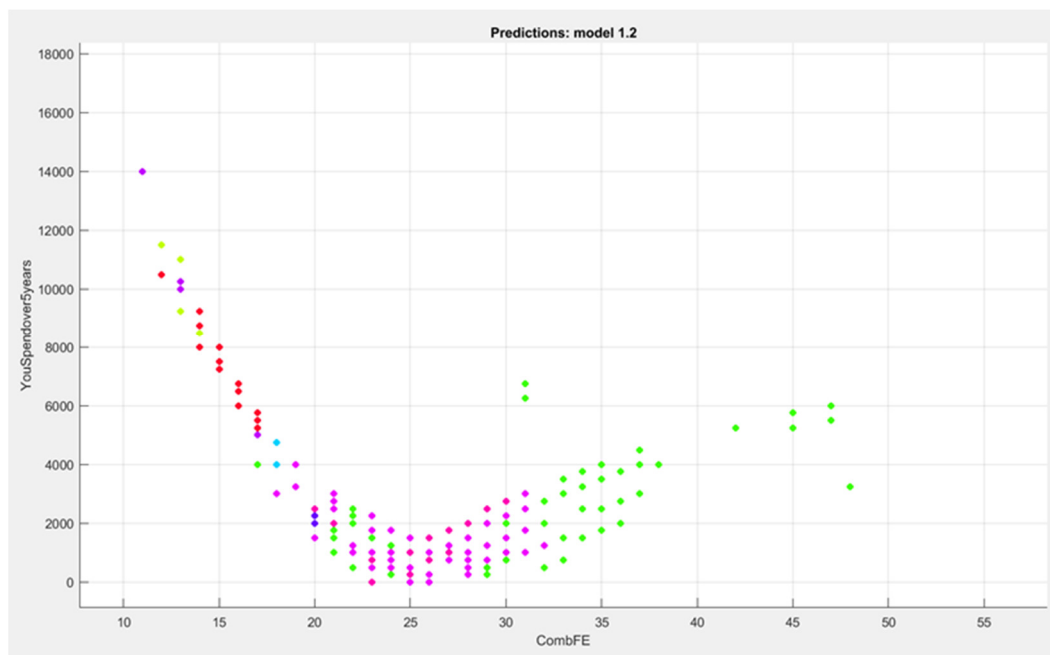


Figure 5-13 Scatter plot for the correct instances using the decision trees classifiers of variable “spent of last five years of fuel” (measured in \$USD) vs “use of fuel in the city” (measured in miles per gallon). Considered instances: Ferrari, Honda, McLaren Automotive, Pagani Automobili S, Quantum Fuel System, Roush, Subaru, Volkswagen, and Aston Martin.

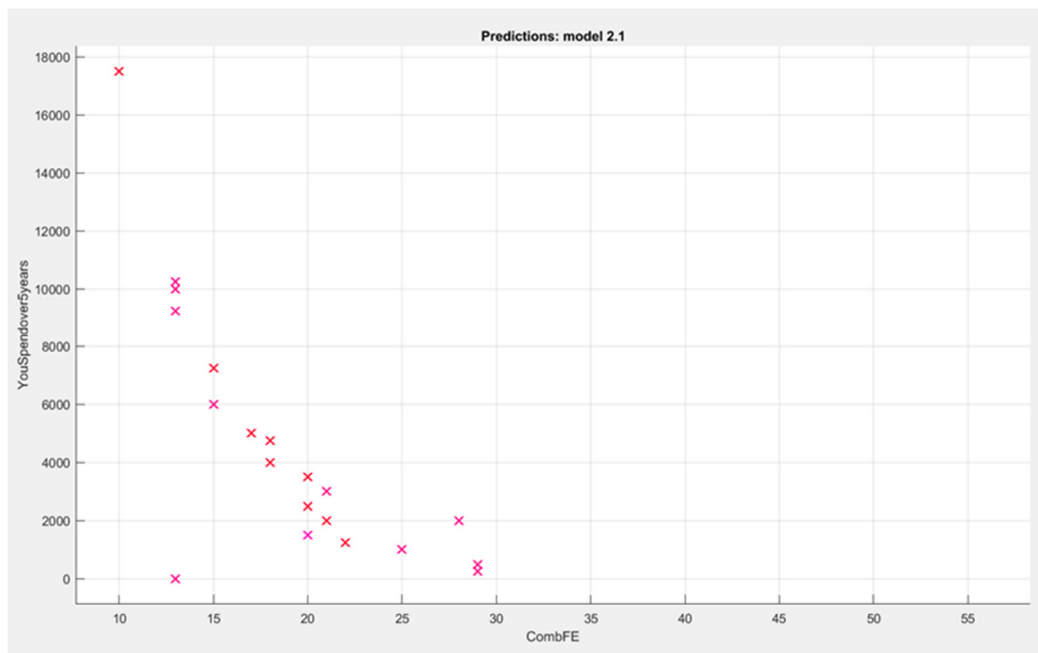


Figure 5-14 Scatter plot for the incorrect instances using the decision trees classifiers of variable “spent of last five years of fuel” (measured in \$USD) vs “use of fuel in the city” (measured in miles per gallon).

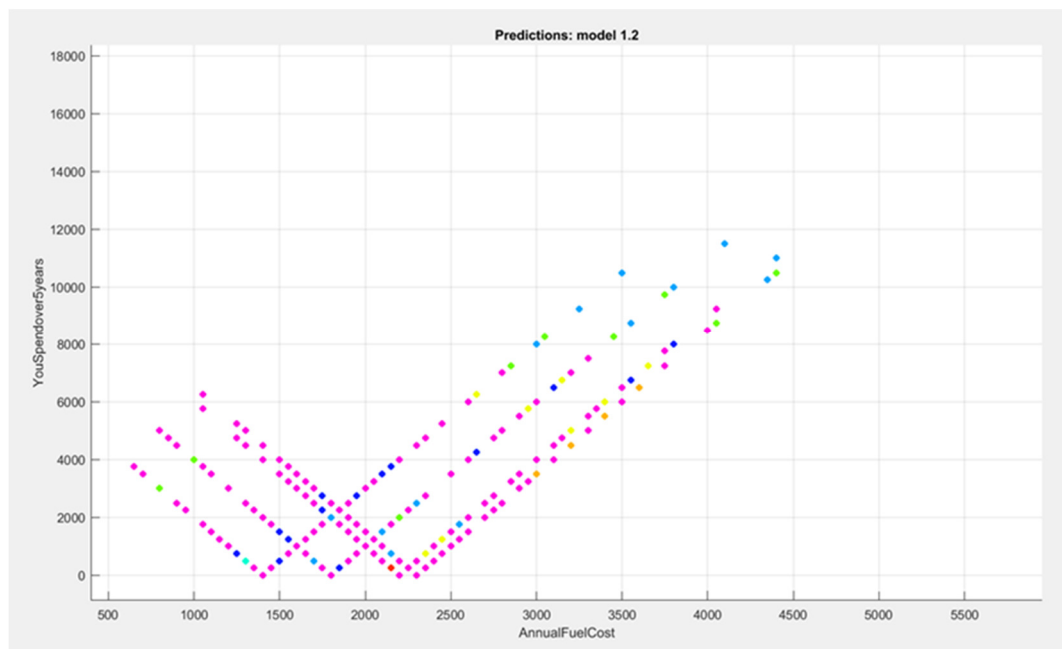


Figure 5-15 Scatter plot for the correct instances using the decision trees classifiers of variable “spent of last five years of fuel” (measured in \$USD) vs “annual cost of

conventional fuel” (measured in \$USD). Considered instances: BMW, Chrysler Group LLC, FCA US LLC, General Motors, Mazda, Mercedes-Benz, Nissan, and Toyota.

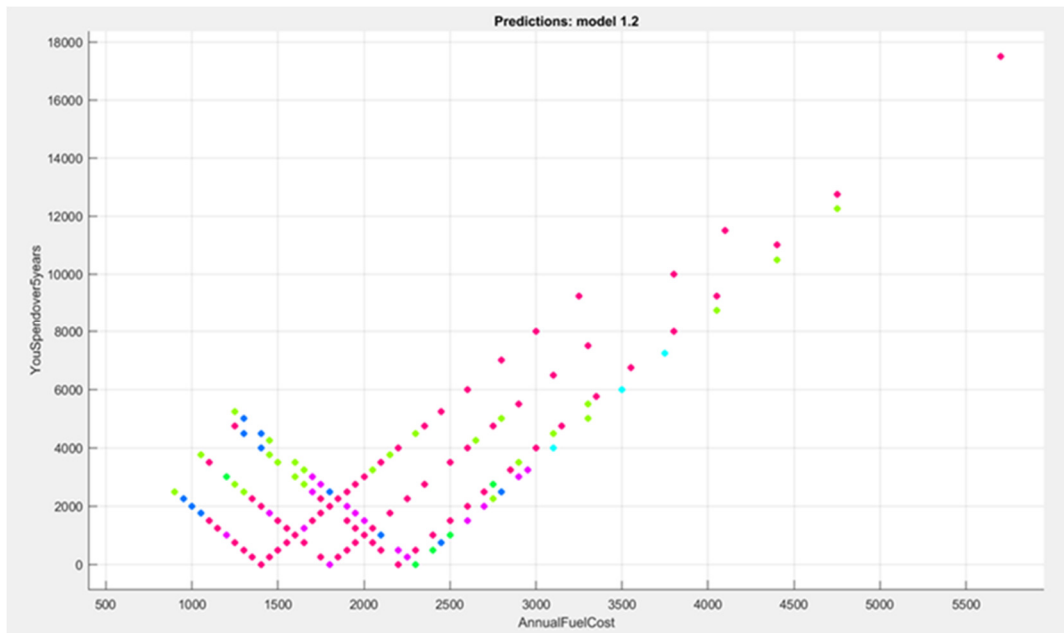


Figure 5-16 Scatter plot for the correct instances using the decision trees classifiers of variable “spent of last five years of fuel” (measured in \$USD) vs “annual cost of conventional fuel” (measured in \$USD). Considered instances: Ford Motor, KIA, Maserati, Mitsubishi Motors Co, Subaru, and Volkswagen Group.

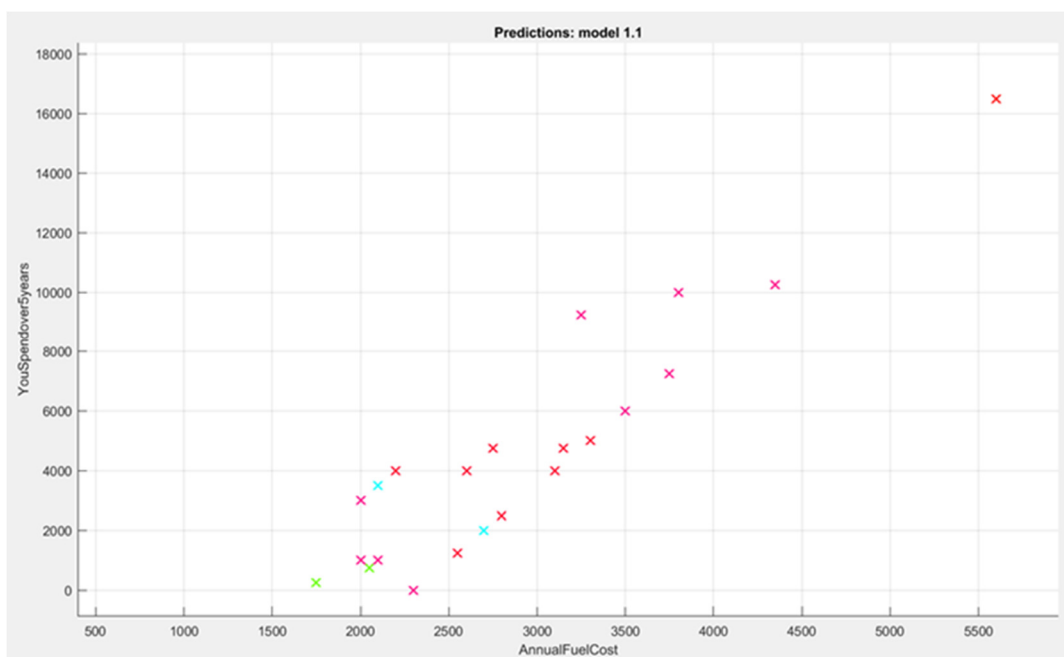


Figure 5-17 Scatter plot for the incorrect instances using the decision trees classifiers of variable “spent of last five years of fuel” (measured in \$USD) vs “annual cost” (measured in \$USD). Considered instances: Audi, General Motors, Maserati, Volkswagen Group, and Aston Martin.

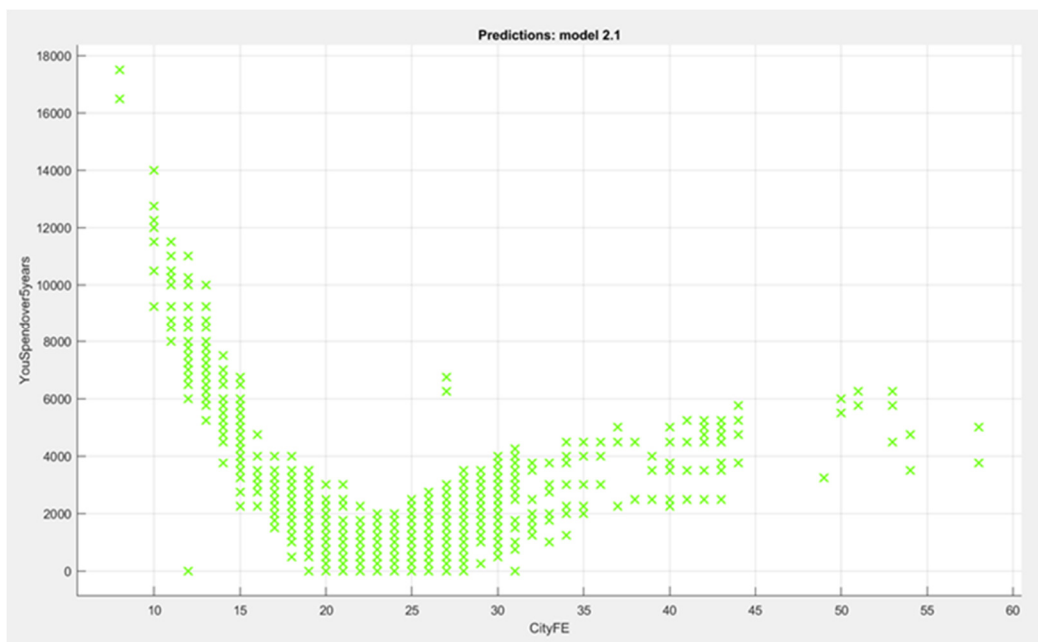


Figure 5-18 Scatter plot for the incorrect instances using the SVM classifiers of variable “spent of last five years of fuel” (measured in \$USD) vs “the use of fuel in the city” (measured in miles per gallon).

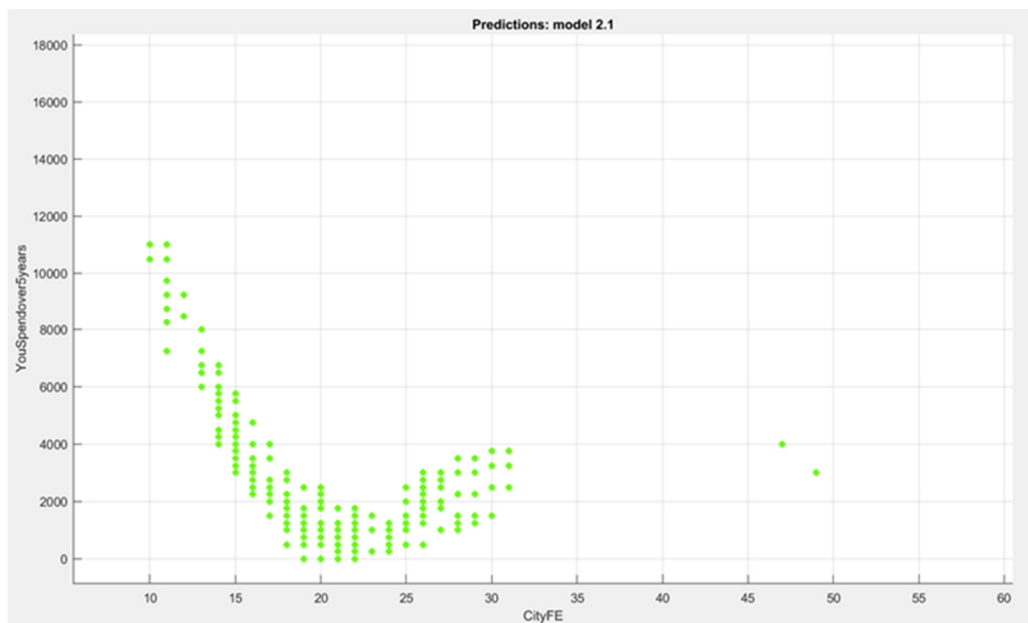


Figure 5-19 Scatter plot for the correct instances using the SVM classifiers of variable “spent of last five years of fuel” (measured in \$USD) vs “the use of fuel in the city” (measured in miles per gallon).

After obtaining the scatter plots for the predictive models, it was necessary to assess the classifier performance, in which a confusion matrix was used to understand how the currently selected classifiers obtained the desired performance in each class. The confusion matrix helps to identify the areas where classification was performed poorly. On the plot depicted in Figure 5-20, each row shows the true class, and the columns depict predictive class. Diagonally, each cell shows where the true class matched with the predictive class. Cells coloured green indicate that the classifier performed well, and observations of this true class were correct. Cells coloured red indicate that the classifier worked poorly, and there was no significance of this predictor in the model. The obtained results for decision trees classifier bagged trees, and SVM are presented in Figure 5-20, Figure 5-21, and Figure 5-22.

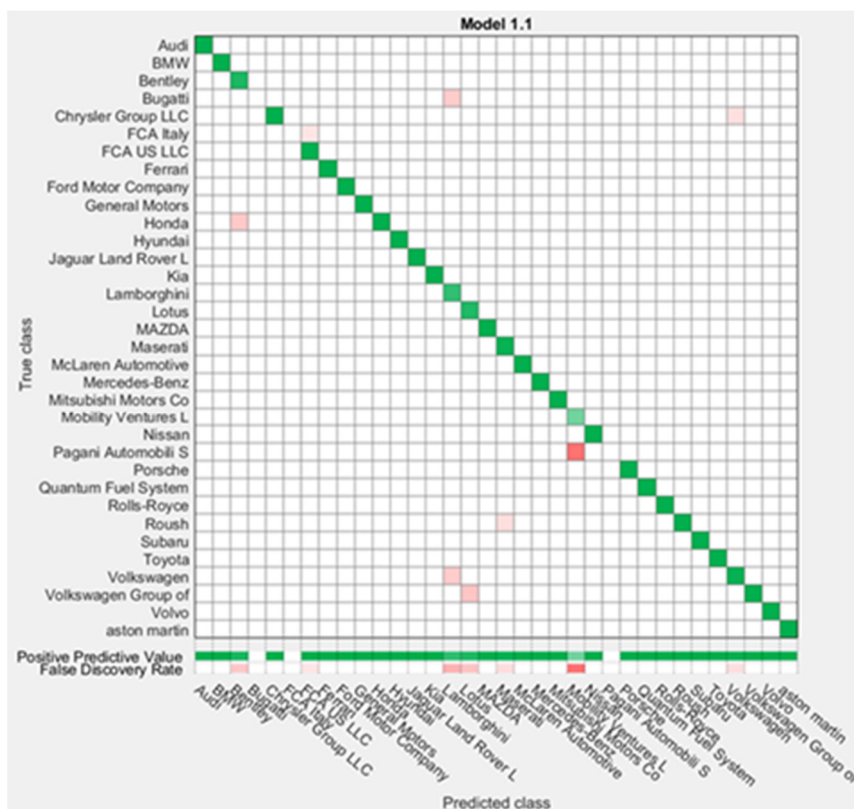


Figure 5-20 Confusion matrix for decision tree classifier showing true class vs. predictive class.

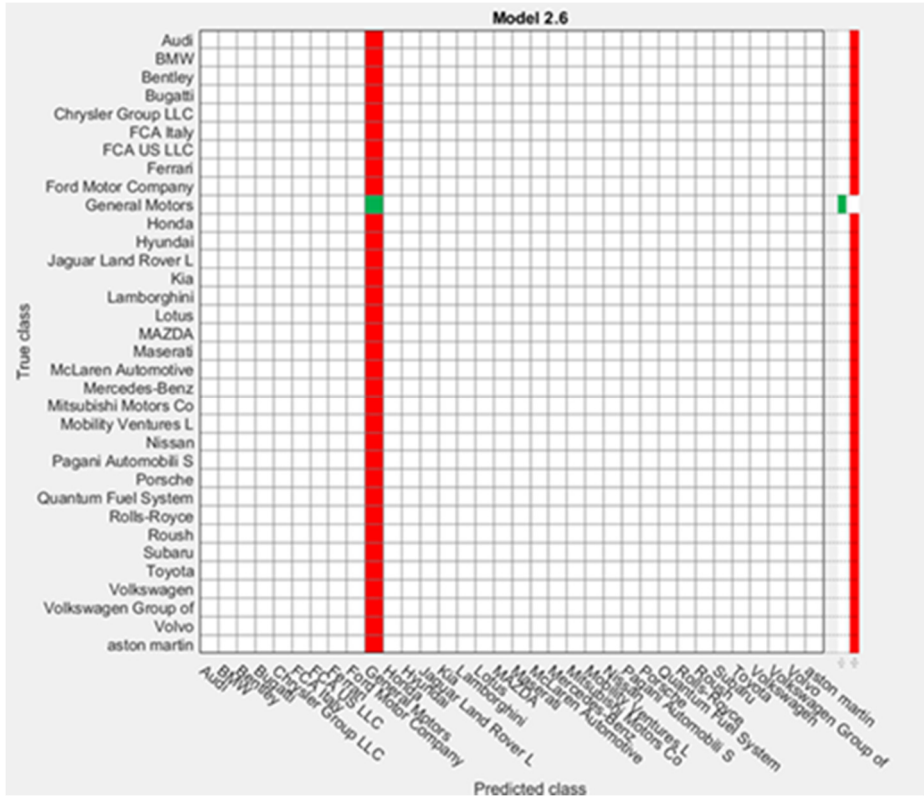


Figure 5-21 Confusion matrix for SVM classifier showing true class vs. predictive class.

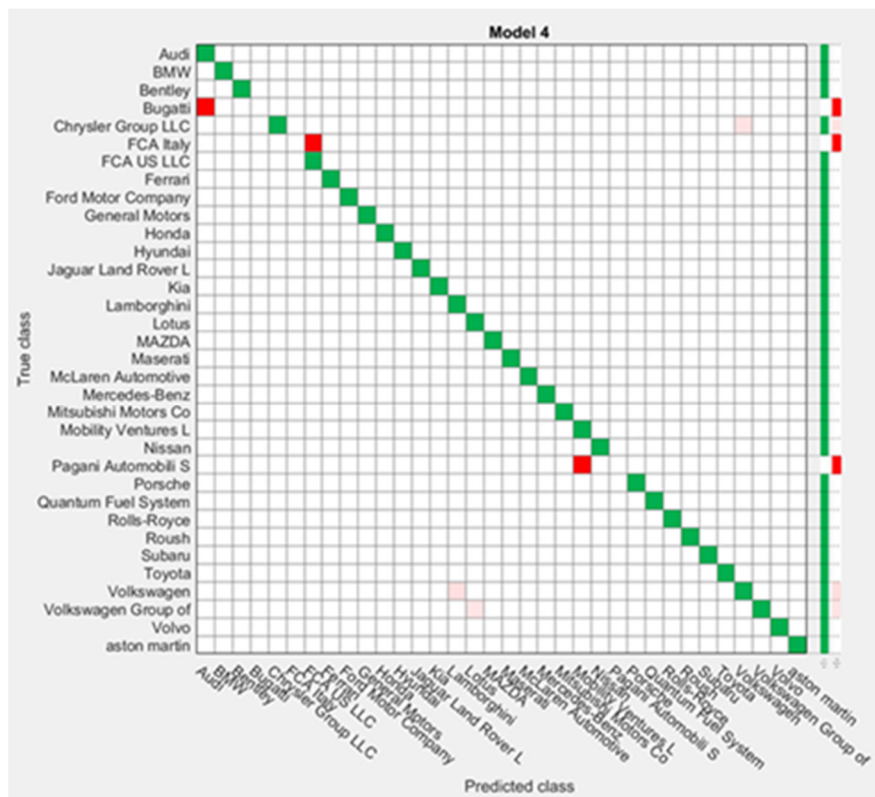


Figure 5-22 Confusion matrix for bagged decision trees classifier showing true class vs predictive class.

Since everything was running in parallel, to determine significant features to include or exclude in the predictive model, we used the parallel coordinates plot. Patterns are displayed in a 2-dimensional plot but correspond to high-dimensional data. Here the selection could be identified, but it also helps to understand relationships between features and useful predictors for separating classes. The training data was utilized, and misclassified points are depicted as dashed lines in Figure 5-23 and Figure 5-24. Figure 5-23 presents the plot that corresponds to the categorical instances of the fuel economy dataset. The standardized values are used to see the distribution of the predictors (manufacturers' name) along the mean distribution on the interaction between each feature, for the figure mentioned above.

It is found that predictors such as Volkswagen and Volvo presented a distribution along the mean for correctly classified instances. For the relationship between the variable,

city unadjusted fuel economy and the predictor, the distributions were outside the mean. Therefore, the variable, city unadjusted fuel economy is less significant for a predictive model. Moving forward with this plot, in the case of Aston Martin predictor, the plot depicts a dotted line from the centre to the right, meaning that incorrect classification was found, and this predictor could definitely be excluded from the model. Regarding model prediction, the significant interactions between variables are “fuel usage”, “annual fuel consumption”, “spend over the last five years of fuel”, and “CO₂ emissions”.

In Figure 5-24, the parallel coordinates plot for numerical instances using normalized values is presented. This figure shows the normalized values or normal distribution of the data, for which the variable, city unadjusted fuel economy (FE) spent on conventional fuel, is significant for the predictive model. According to the information provided in [126], the rates of city unadjusted FE spent on conventional fuel variable, describes the consumption of unadjusted conventional fuel, for single-fuel vehicles.

The other significant interaction between variables is \$spent over 5 years vs the predictors. The data collected for this variable reflects how much users of vehicles spent over the last 5 years compared to average cars and the information provided by each manufacturer. The instances Lamborghini, Aston Martin, and Volkswagen represent the largest fuel expense and therefore not considered for the predictive model. Similar to the plot of categorical variables, the instances Lamborghini and Aston Martin presented misclassification. Once these instances (Aston Martin and Lamborghini) were not considered in the model, resulted in a better prediction result.

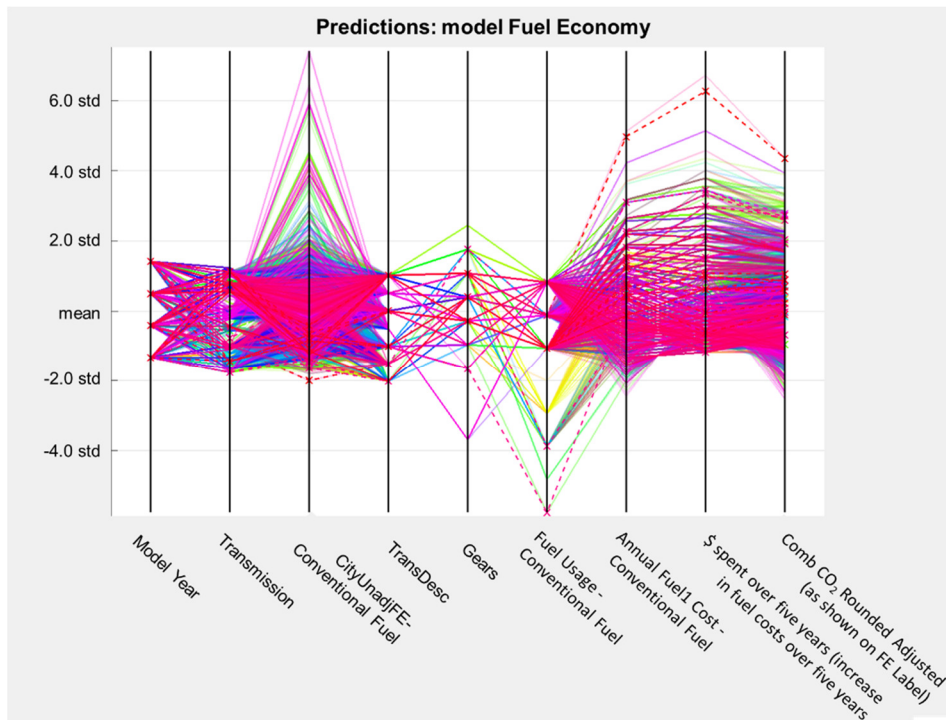


Figure 5-23 Standardized values used for the parallel coordinates plot of categorical instances of the fuel economy data for selection of features.

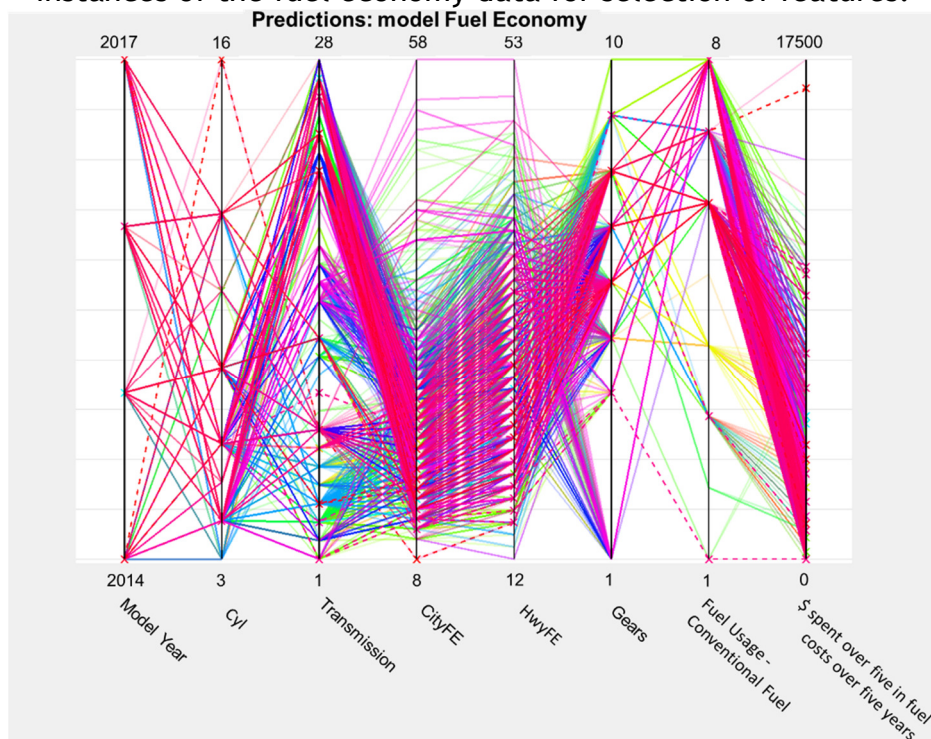
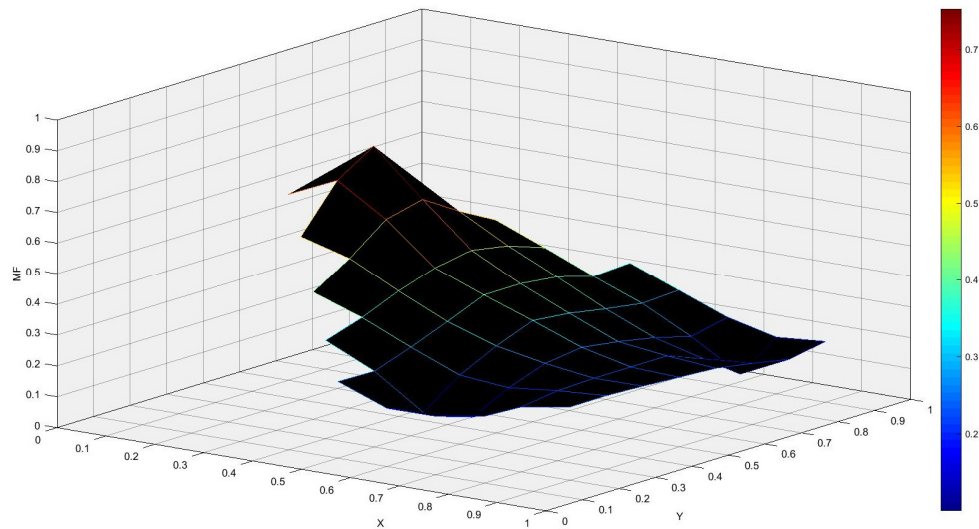


Figure 5-24 Normalized values used for the parallel coordinates plot of categorical instances of the fuel economy data for selection of features.

Fuzzy c-means Clustering

Cluster analysis was performed by MATLAB, using the fuzzy logic toolbox for pattern identification. The fuzzy c-means updated the cluster centres and membership grades of each data point. The clusters obtained were iteratively moved from the centre rightward in the dataset. The selected parameters for the fuzzy c-means were 3 clusters, exponent = 3, the maximum of iterations = 100, and minimum improvement = 1e-05. For this objective function, the iterations are based on minimizing an objective function that represents the distance from any given data point to a cluster centre weighted by that data point's membership grade.

The obtained membership function plots are presented in Figure 5-25. It is shown in this plot the times that each cluster reached the maximum of iterations, or when the objective function improvement between two consecutive iterations is less than the minimum amount of improvement specified. Figure 5-26 shows the identification and partition of 3 clusters, which represent each membership function.



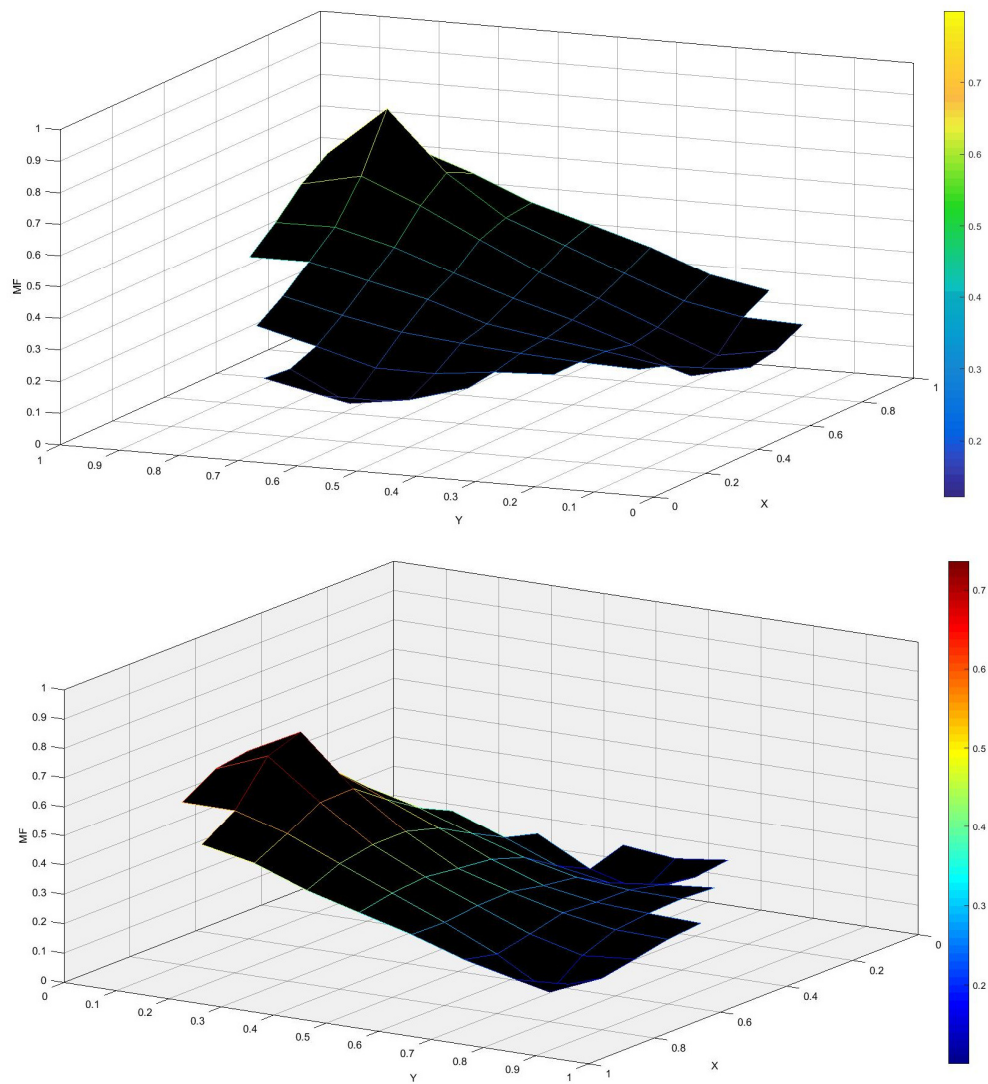


Figure 5-25 Membership function plots for the fuzzy c-means clustering. From top to bottom: cluster 1, 2, and 3.

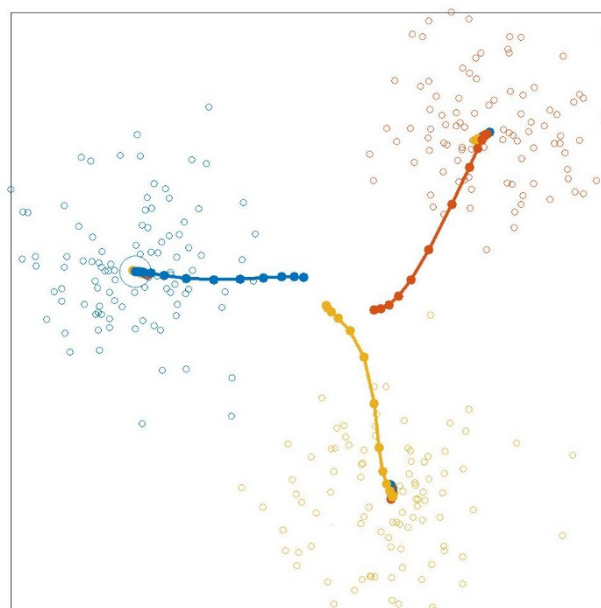


Figure 5-26 Fuzzy c-means partition of 3 clusters plot.

Fuzzy clustering was proved to be helpful to demonstrate the framework proposed in chapter 4, section 3 as the classification learner approach for i4 environments. As stated in chapter 3, section 1 the fuzzy cluster were applied successfully to a mix of categorical and numerical inputs, and the instances were split into 3 clusters, and the membership function plots in Figure 5-25 show the degree of belonging to different clusters, represented by each membership function. In such analysis, it can be inferred that cluster 2 has a crisper degree of membership, noticed by the peak in the plot reaching more than 0.7 degrees of membership.

5.3.4 CPU Dataset Results

The results of the CPU dataset analysis are presented in this section, where we used the Matlab classification learner app to train the CPUs dataset. To obtain the classification model, as mentioned before, product collection variable was selected as a response, and all the other variables were considered predictors. Matlab classification learner was mainly used because of the automation feature that enables us to run several parallel classifiers and see which can obtain the best predictive model, also because of the process-ability of importing the raw data without investing too much time making adjustments. Since there were some attributes that shown no entries or values, we remove those from the trained dataset to give a better adjustment to the classification model. Different to the analysis presented in previous sections of this chapter, the CPU dataset represented a more complete challenge and is addressed by the framework proposed in Figure 4-4 from chapter 4, section 4; in which we implemented a full analysis, including the statistical analysis as a way of validation for the obtained models. Feature selection is now implemented and added to the closed-loop cycle to complete the full automation of bigger datasets since in previous cases (applications) was not required to perform a complete automated data mining analysis.

This dataset used to train the classification model encompassed 39 attributes split into 9 numerical values, and 30 categorical. The total observations considered in this dataset were 2298.

Attribute Classification

In Figure 5-27 to Figure 5-32, the scatter plots showing model predictors, and the correct and incorrect instances for the CPUs dataset are depicted, as well as interactions between variables. Figure 5-27 presents the corresponding instances classified correctly for the ensemble bagged tree classifier, the colours presented in this figure correspond to each product collection name (response), and the identification of colours vs product name will be presented in detail below. In this figure is presented the interaction between recommended price vs processor number using the correct model predictor instances. In Figure 5-28, similar to the previous figure, the incorrect classified observations are depicted. The incorrect instances presented in this plot were: Intel Celeron® Processor 1000 Series, Legacy Intel Core Processors, Legacy Intel® Pentium® Processor, and Legacy Intel® Xeon® Processors.









































For better visualization purposes we decided to break down same classification interaction scatter plots using different product collection names (response instance). Following with the identification of classification instances shown in scatter plots, in Figure 5-29 is presented the correct model predictors for the interaction of recommended customer price variable and number of cores, which include: Intel® Atom Processor C Series, Intel Itanium® Processor 9100 Series, Intel® Xeon Phi x200 Product Family, Intel® Xeon® Processor D Family, Intel® Xeon® Processor E3 v3 Family, Intel® Xeon® Processor E5 Family, Intel® Xeon® Processor E5 v2 Family, Intel® Xeon® Processor E5 v3 Family, Intel® Xeon® Processor E5 v4 Family, Intel® Xeon® Processor E7 Family, Intel® Xeon® Processor E7 v2 Family, Intel® Xeon® Processor E7 v3 Family, Intel® Xeon® Processor E7 v4 Family, Intel® Xeon® Processor W Family, Intel® Xeon® Scalable Processors, Legacy Intel® Celeron® Processor, Legacy Intel® Core Processors, Legacy Intel® Pentium® Processor, and Legacy Intel® Xeon® Processors identified as true predictors. The incorrect classified instances are depicted in Figure 5-30 for the interaction between recommended customer price vs number of cores variable, and it was encountered the following product collection names: 5th Generation Intel® Core i5 Processors, 7th Generation Intel® Core i3 Processors, 7th Generation Intel® Core i3 Processors, Legacy Intel® Celeron® Processor, Legacy Intel® Core Processors, Legacy Intel® Pentium® Processor, and Legacy Intel® Xeon® Processors.

In Figure 5-31 is shown the scatter plot for the correct instances classification of variables recommended customer price and temperature, in which all the classes were considered to be depicted. The misclassified variables (recommended customer price and temperature) are presented in Figure 5-32, where it was found the following classes to have an impact on this interaction: 4th Generation Intel® Core i5 Processors, Intel® Xeon® Processor E3 v3 Family, Intel® Xeon® Processor E5 v2 Family, Legacy Intel® Core Processors, and Legacy Intel® Xeon® Processors.

The above-mentioned scatter plots help to investigate patterns, features, and how the product collection (response) prediction performs against the selected predictors (all the other variables).

For the scatter plots, the following range of colours was used to identify each class or the product collection variable, each name corresponds to different processors contained in the dataset:

■ 4th Generation Intel® Core™ i3 Processors, ■ 4th Generation Intel® Core™ i5 Processors, ■ 4th Generation Intel® Core™ i7 Processors, ■ 5th Generation Intel® Core™ M Processors, ■ 5th Generation Intel® Core™ i3 Processors, ■ 5th Generation Intel® Core™ i5 Processors, ■ 5th Generation Intel® Core™ i7 Processors, ■ 6th Generation Intel® Core™ i3 Processors, ■ 6th Generation Intel® Core™ i5 Processors, ■ 6th Generation Intel® Core™ i7 Processors, ■ 6th Generation Intel® Core™ m Processors, ■ 7th Generation Intel® Core™ i3 Processors, ■ 7th Generation Intel® Core™ i5 Processors, ■ 7th Generation Intel® Core™ i7 Processors, ■ 7th Generation Intel® Core™ m Processors, ■ 8th Generation Intel® Core™ i5 Processors, ■ 8th Generation Intel® Core™ i7 Processors, ■ Intel® Atom™ Processor C Series, ■ Intel® Atom™ Processor D Series, ■ Intel® Atom™ Processor E Series, ■ Intel® Atom™ Processor N Series, ■ Intel® Atom™ Processor S Series, ■ Intel® Atom™ Processor X Series, ■ Intel® Atom™ Processor Z Series, ■ Intel® Celeron® Processor 1000 Series, ■ Intel® Celeron® Processor 2000 Series, ■ Intel® Celeron® Processor 3000 Series, ■ Intel® Celeron® Processor G Series, ■ Intel® Celeron® Processor J Series, ■ Intel® Celeron® Processor N Series, ■ Intel® Core™ X-series Processors, ■ Intel® Itanium® Processor 9000 Series, ■ Intel® Itanium® Processor 9100 Series, ■ Intel® Itanium® Processor 9300 Series, ■ Intel® Itanium®

Processor 9500 Series,  Intel® Itanium® Processor 9700 Series,  Intel® Itanium® Processors with 400 MHz FSB,  Intel® Itanium® Processors with 533 MHz FSB,  Intel® Itanium® Processors with 677 MHz FSB,  Intel® Pentium® Processor 1000 Series,  Intel® Pentium® Processor 2000 Series,  Intel® Pentium® Processor 3000 Series,  Intel® Pentium® Processor 4000 Series,  Intel® Pentium® Processor D Series,  Intel® Pentium® Processor G Series,  Intel® Pentium® Processor J Series,  Intel® Pentium® Processor N Series,  Intel® Quark®,[®] Microcontroller D1000 Series,  Intel® Quark®,[®] Microcontroller D2000 Series,  Intel® Quark®,[®] SE C1000 Microcontroller Series,  Intel® Quark®,[®] SoC X1000 Series,  Intel® Xeon Phi®,[®] x100 Product Family,  Intel® Xeon Phi®,[®] x200 Product Family,  Intel® Xeon® Processor D Family,  Intel® Xeon® Processor E3 Family,  Intel® Xeon® Processor E3 v2 Family,  Intel® Xeon® Processor E3 v3 Family,  Intel® Xeon® Processor E3 v4 Family,  Intel® Xeon® Processor E3 v5 Family,  Intel® Xeon® Processor E3 v6 Family,  Intel® Xeon® Processor E5 Family,  Intel® Xeon® Processor E5 v2 Family,  Intel® Xeon® Processor E5 v3 Family,  Intel® Xeon® Processor E5 v4 Family,  Intel® Xeon® Processor E7 Family,  Intel® Xeon® Processor E7 v2 Family,  Intel® Xeon® Processor E7 v3 Family,  Intel® Xeon® Processor E7 v4 Family,  Intel® Xeon® Processor W Family,  Intel® Xeon® Scalable Processors,  Legacy Intel Atom® Processors,  Legacy Intel® Celeron® Processor,  Legacy Intel® Core®,[®] Processors,  Legacy Intel® Pentium® Processor, and  Legacy Intel® Xeon® Processors.

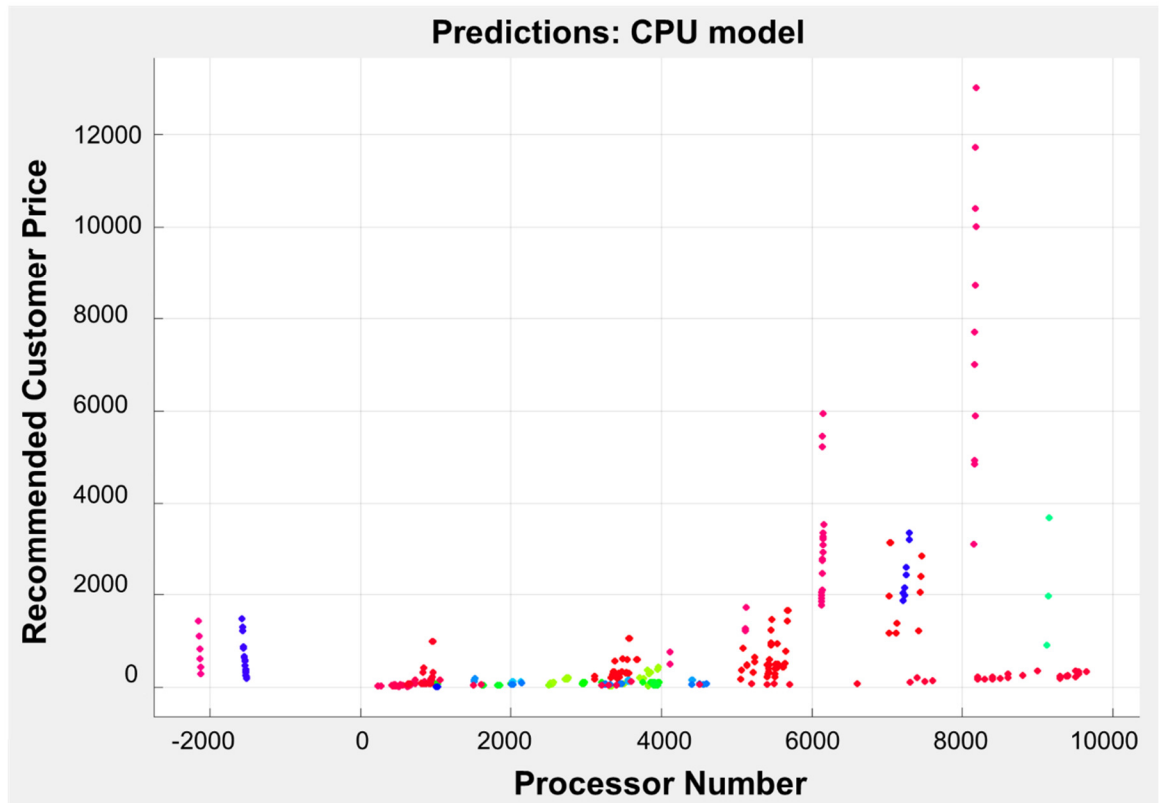


Figure 5-27 Scatter plot for the correct instances using ensemble bagged trees classifier of variable “recommended customer price” (measured in \$USD) vs “processor number” (unit number). All classes included.

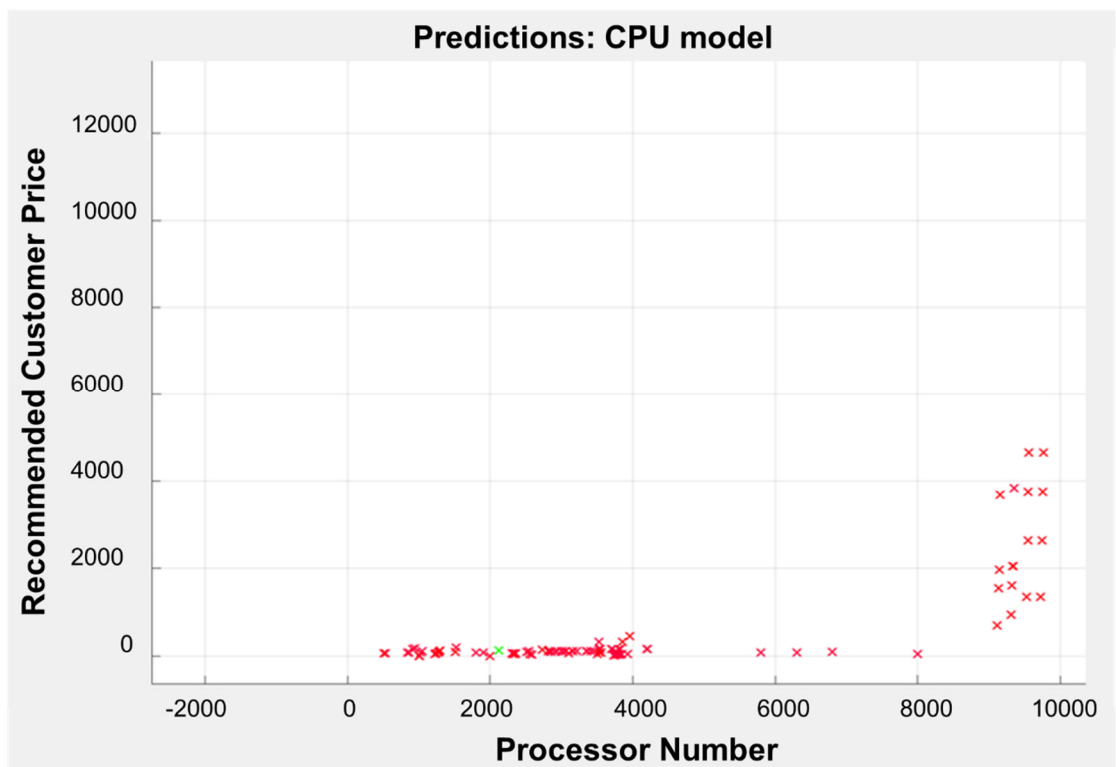


Figure 5-28 Scatter plot for the incorrect instances using ensemble bagged tree classifier of variable “recommended customer price” (measured in \$USD) vs “processor number” (unit number). Considered instances: Intel Celeron®

Processor 1000 Series, Legacy Intel Core Processors, Legacy Intel® Pentium® Processor, and Legacy Intel® Xeon® Processors.

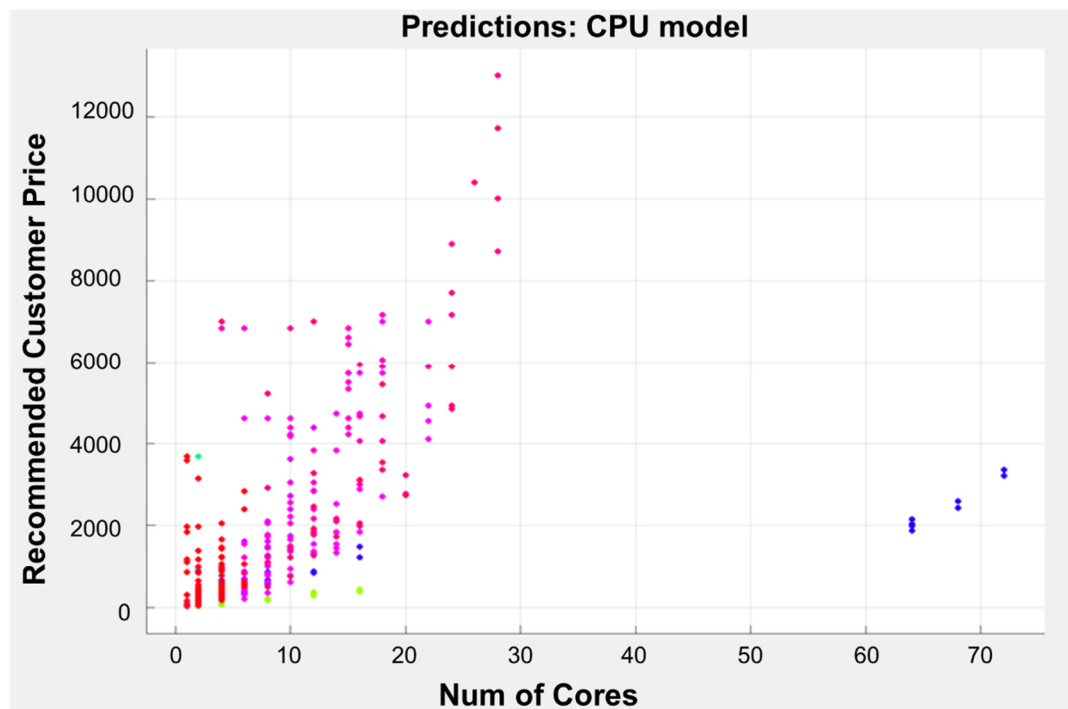


Figure 5-29 Scatter plot for the correct instances using ensemble bagged trees classifiers of variable “recommended customer price” (measured in \$USD) vs “number of cores” (unit). Considered instances: Intel® Atom Processor C Series, Intel Itanium® Processor 9100 Series, Intel® Xeon Phi x200 Product Family, Intel® Xeon® Processor D Family, Intel® Xeon® Processor E3 v3 Family, Intel® Xeon® Processor E5 Family, Intel® Xeon® Processor E5 v2 Family, Intel® Xeon® Processor E5 v3 Family, Intel® Xeon® Processor E5 v4 Family, Intel® Xeon® Processor E7 Family, Intel® Xeon® Processor E7 v2 Family, Intel® Xeon® Processor E7 v3 Family, Intel® Xeon® Processor E7 v4 Family, Intel® Xeon® Processor W Family, Intel® Xeon® Scalable Processors, Legacy Intel® Celeron® Processor, Legacy Intel® Core Processors, Legacy Intel® Pentium® Processor, and Legacy Intel® Xeon® Processors.

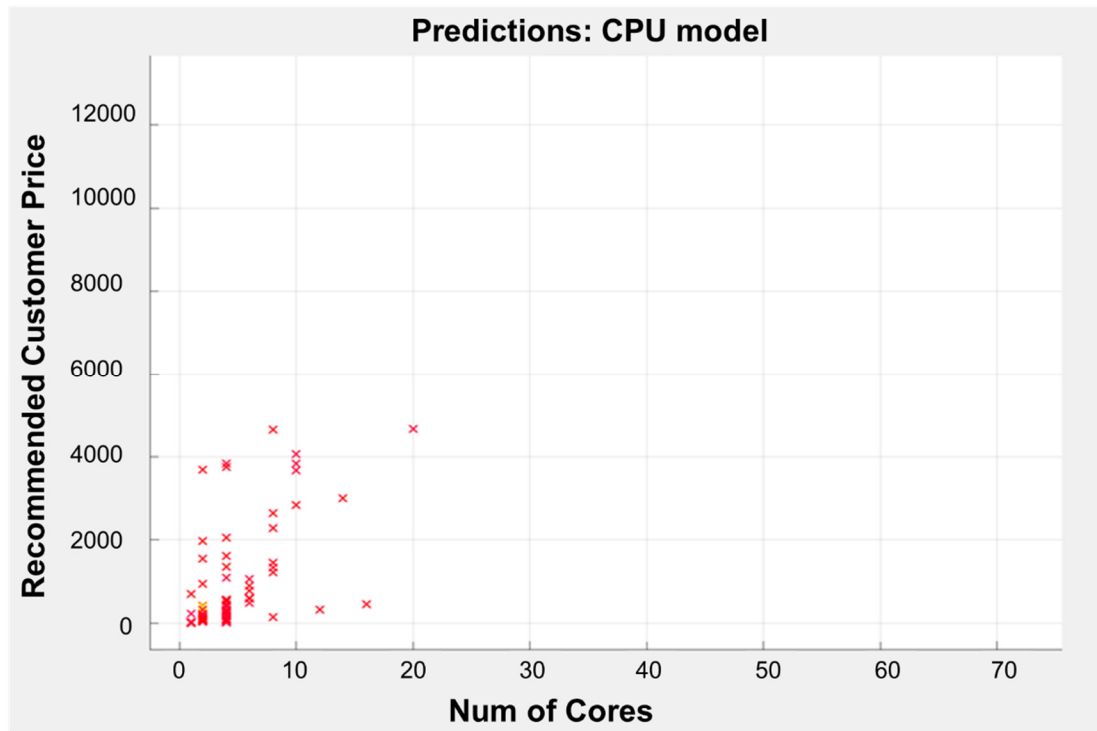


Figure 5-30 Scatter plot for the incorrect instances using ensemble bagged trees classifiers of variable “recommended customer price” (measured in \$USD) vs “number of cores” (unit). Considered instances: 5th Generation Intel® Core i5 Processors, 7th Generation Intel® Core i3 Processors, 7th Generation Intel® Core i3 Processors, Legacy Intel® Celeron® Processor, Legacy Intel® Core Processors, Legacy Intel® Pentium® Processor, and Legacy Intel® Xeon® Processors.

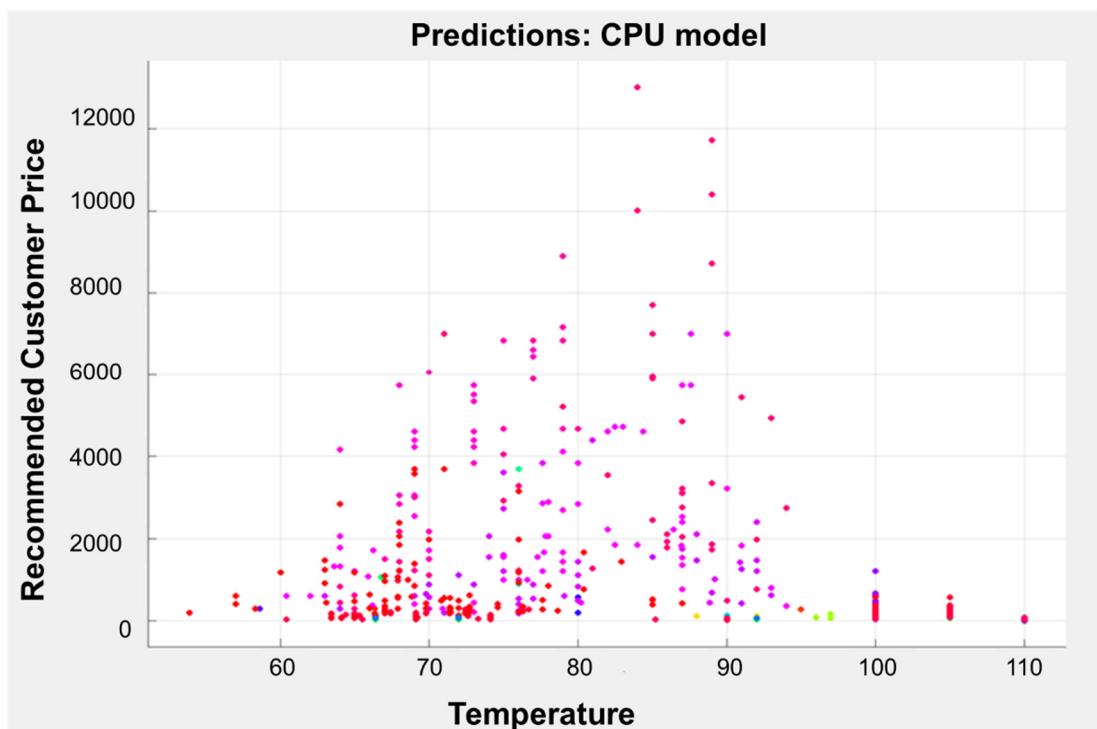


Figure 5-31 Scatter plot for the correct instances using ensemble bagged trees classifiers of variable “recommended customer price” (measured in \$USD) vs “temperature” (C°). All classes included.

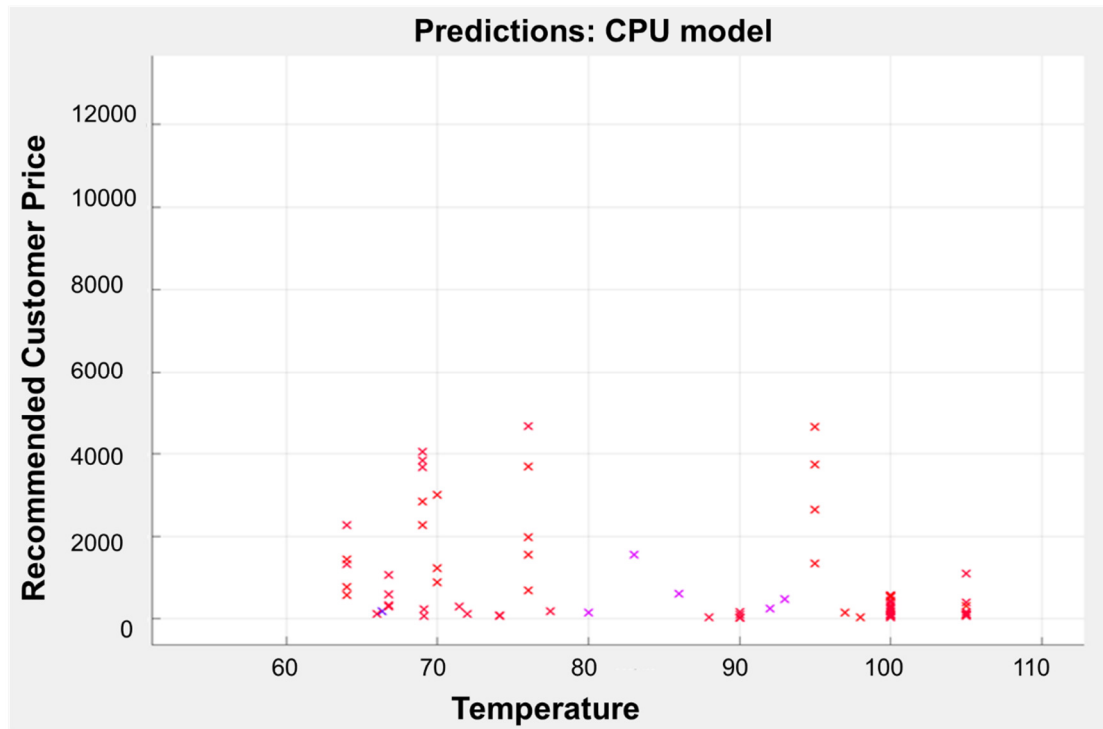


Figure 5-32 Scatter plot for the incorrect instances using ensemble bagged trees classifiers of variable “recommended customer price” (measured in \$USD) vs “temperature” (C°). Considered instances: 4th Generation Intel® Core i5 Processors, Intel® Xeon® Processor E3 v3 Family, Intel® Xeon® Processor E5 v2 Family, Legacy Intel® Core Processors, and Legacy Intel® Xeon® Processors.

After analysing the scatter plots for the predictive models, it was necessary to assess the classifier performance, in which a confusion matrix was used to understand how the currently selected classifiers obtained the desired performance in each class. The confusion matrix helps to identify the areas where classification was performed poorly. On the plot depicted in Figure 5-33, each row shows the true class, and the columns depict predictive class. Diagonally, each cell shows where the true class matched with the predictive class. Cells coloured green indicate that the classifier performed well, and observations of this true class were correct. Cells coloured red indicate that the classifier worked poorly, and there was no significance of this predictor in the model.

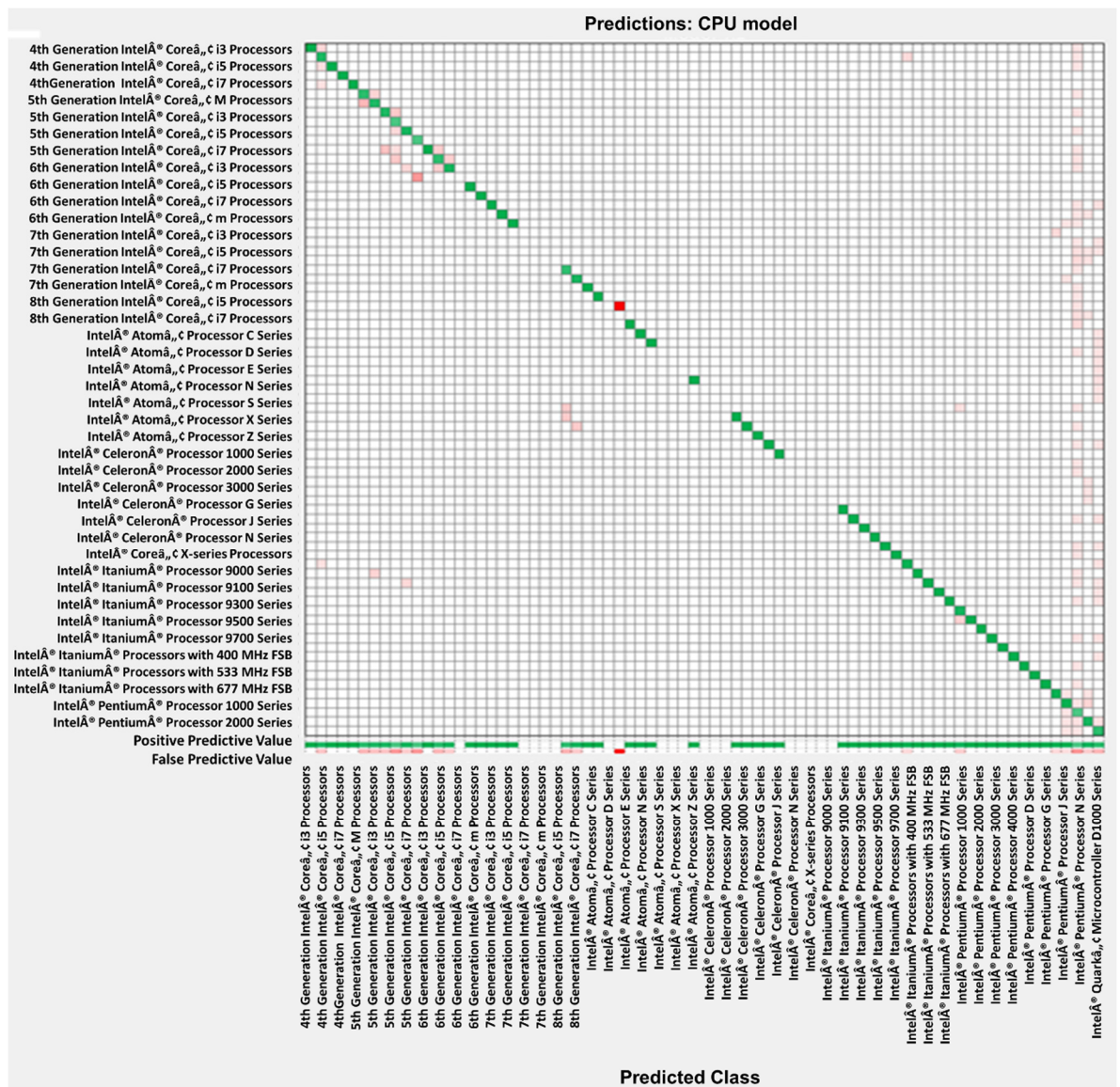


Figure 5-33 Confusion matrix for the ensemble bagged tree classifier showing true class vs. predictive class.

Moving forward with the analysis, one of the benefits of training the dataset in parallel is to determine significant features to include or exclude in the predictive model, using the parallel coordinates plot. Patterns are displayed in a 2-dimensional plot but correspond to high-dimensional data. Here the selection could be identified, but it also helps to understand relationships between features and useful predictors for separating classes. The training data was utilized, and misclassified points are depicted as dashed lines in Figure 5-34. The standardized values are used to see the distribution of the predictors along the mean distribution of the interaction between each feature, for the figure mentioned above. We found that predictors such as vertical segment, recommended customer price, thermal design power, max memory size, temperature, and memory type presented a distribution along the mean for correctly classified

instances. For the relationship between the variable number of cores, ECC memory support, max memory bandwidth, and the response, the distributions were outside the mean and showing misclassification. Therefore, these variables are less significant for the classification model.

In Figure 5-35, the parallel coordinates plot for numerical instances using normalized values is presented. This figure shows the normalized values or normal distribution of the data, for which the variables recommended customer price, processor number, processor base frequency, bus speed, max memory size are significant predictors for the classification model.

The plot presented in Figure 5-36 helps in a different part of the analysis, that is, which observations inside the response have poor classification rates. The selected observation is Intel® Celeron® Processor J Series, and show a rate of 0 %, determined by the current classifier red dot. This plot refers to the receiver operating characteristic (ROC) curve that shows true and false positive rates. And the area under the curve measures the overall quality of the classifier.

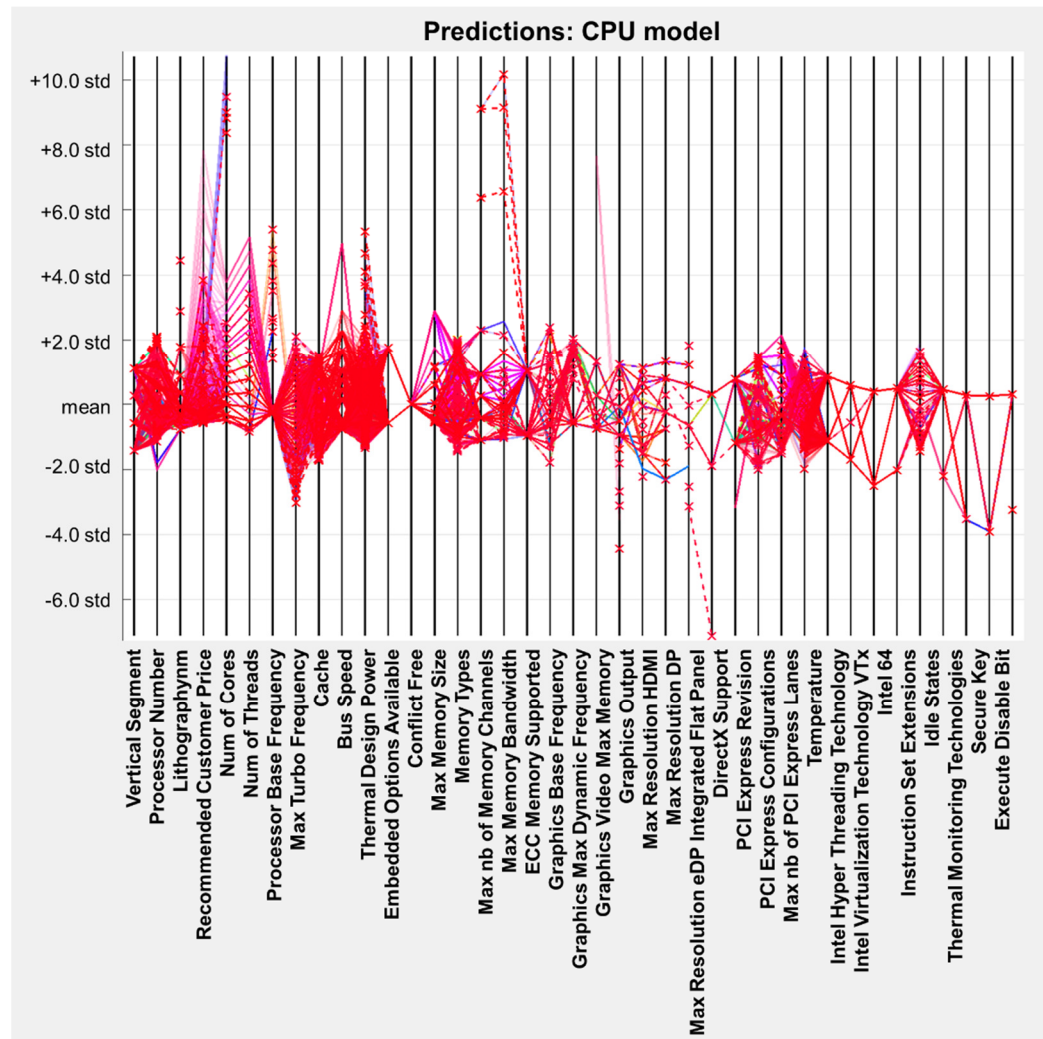


Figure 5-34 Standardized values used for the parallel coordinates plot of categorical instances of the CPUs data for selection of features.

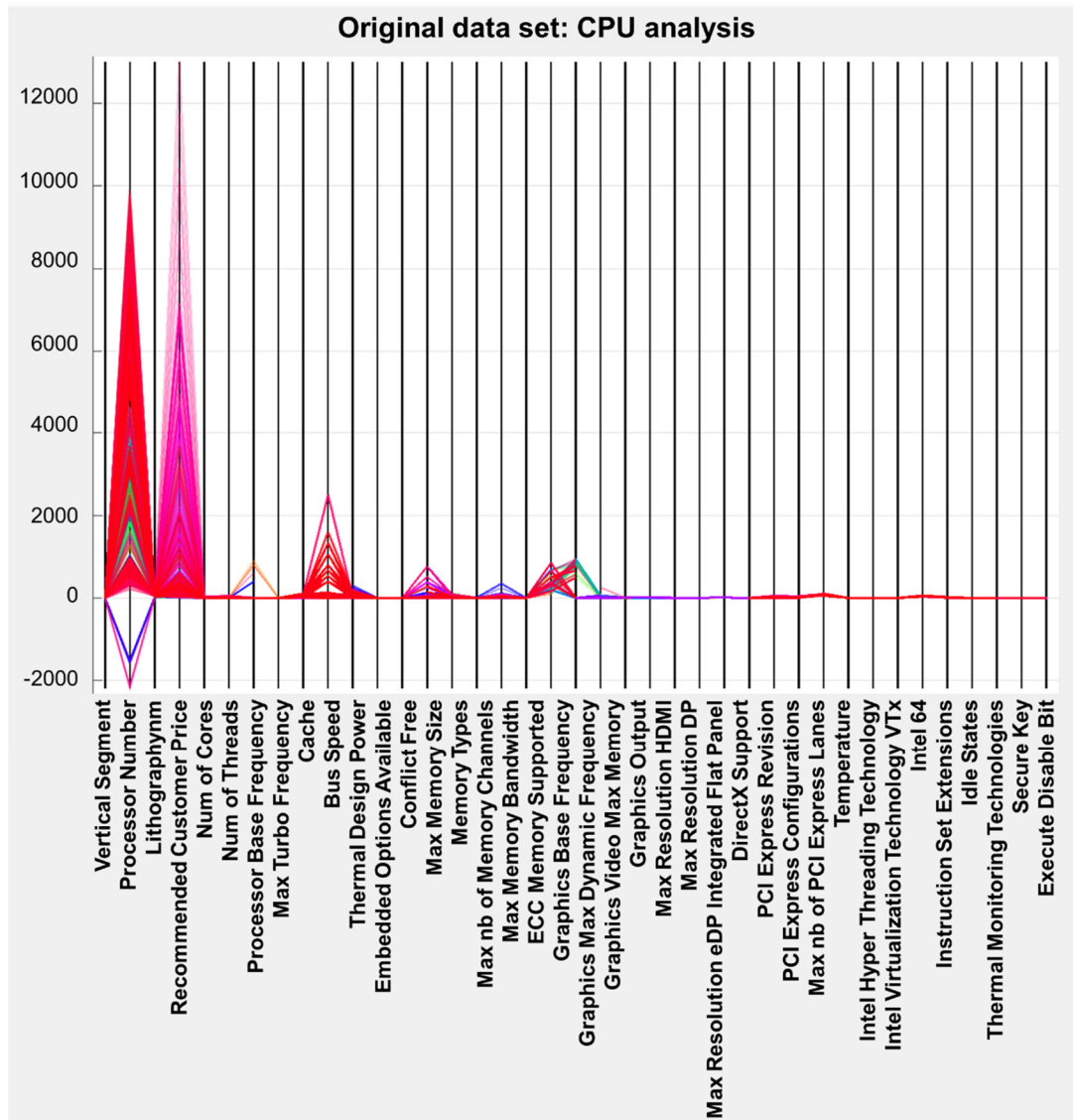


Figure 5-35 Normalized values used for the parallel coordinates plot of categorical instances of the CPUs data for selection of features.

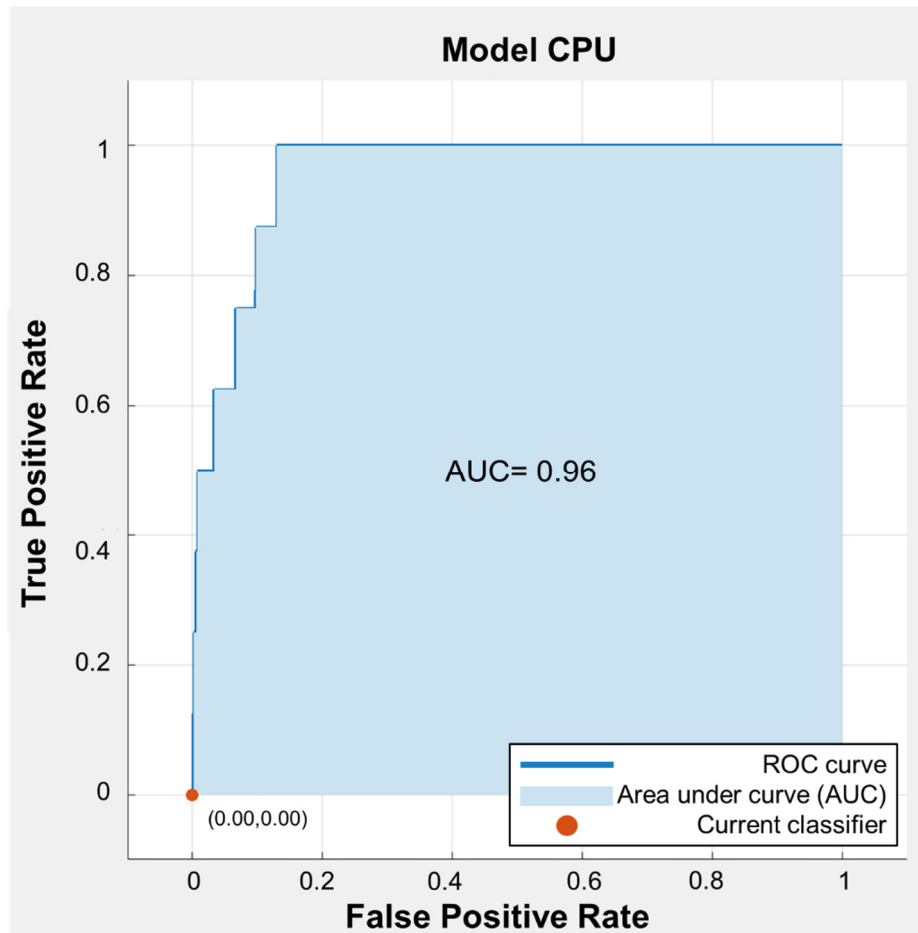


Figure 5-36 ROC curve plot showing the misclassification of the observation Intel® Celeron® Processor J Series.

Feature Selection Using Genetic Search

Once obtained the classification model, and compliant with the accuracy of the trained model, it was performed the second part of the analysis according to the methodology presented in Chapter 4, section 4.4, which is the feature selection analysis. Feature selection was performed using Matlab, combining the classification learner toolbox with a genetic search code for feature selection and clustering, using the code obtained from the trained dataset. The clusters obtained were iteratively moved from the centre rightward in the dataset. Feature selection using genetic search was performed using the following parameters: 1) probability of search = 0.6, 2) maximum of generations = 20, 3) mutation probability = 0.033, and 4) population size = 90. In Figure 5-37 the population growth for the GA using the trained dataset classified previously are presented.

Table 5—10 presents the results obtained from the feature selection analysis using genetic search. Here we present how possible is for an attribute to be selected,

based on how relevant each attribute is for the model. In theory, feature selection can be considered as a combination of search technique to propose a new subset of features (attributes). In this case, GA was used as the evaluator or objective function, each possible subset of attributes was tested, and the percentage shown in Table 5–10 how each feature minimized the error rate is presented.

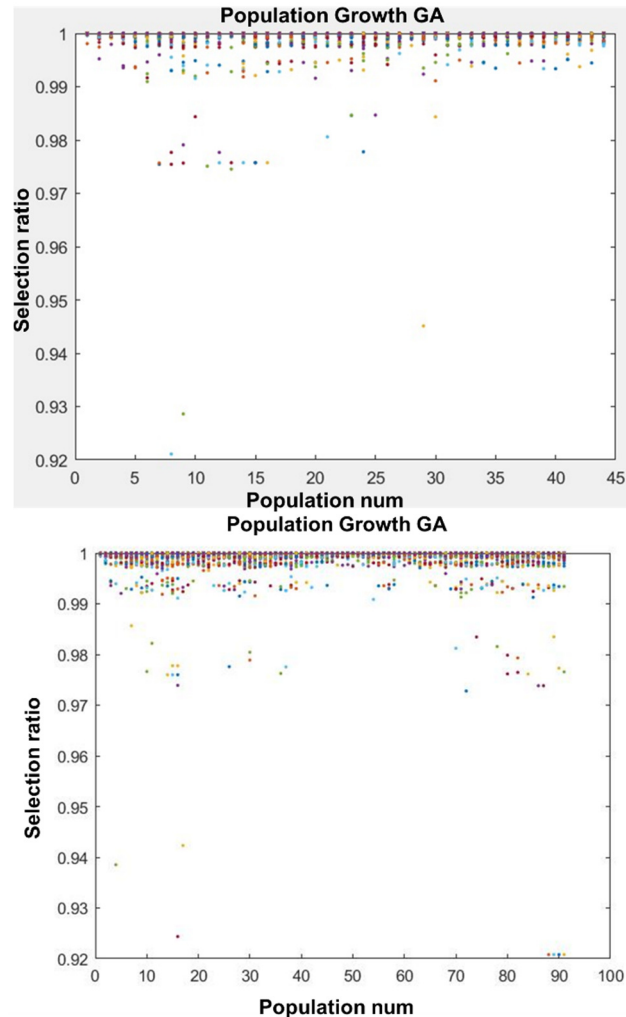


Figure 5-37 Population growth using GA for feature selection.

Table 5–10 Feature selection results using genetic search

Selection %	Order of attribute	Attribute
90	6	Recommended Customer Price
90	13	Thermal Design Power (W)
90	17	Memory Types
90	39	Thermal Monitoring Technologies
80	7	Nb of Cores
80	21	Graphics Base Frequency
80	22	Graphics Max Dynamic Frequency
80	23	Graphics VideoMax Memory
80	27	Max Resolution eDP Integrated Flat Panel
80	29	PCI Express Revision

80	34	Intel Virtualization Technology VTx_
80	36	Instruction Set
80	41	Execute Disable Bit
70	8	Nb of Threads
70	9	Processor Base Frequency
70	10	Max Turbo Frequency
70	19	Max Memory Bandwidth
70	28	DirectX Support
70	31	Max nb of PCI Express Lanes
70	37	Instruction Set Extensions
70	38	Idle States
70	40	Secure Key
60	14	Embedded Options Available
60	18	Max nb of Memory Channels
50	35	Intel 64
40	4	Processor Number
40	16	Max Memory Size
40	24	Graphics Output
30	5	Lithography nm
30	25	Max Resolution HDMI
20	3	Vertical Segment
20	15	Conflict Free
20	20	ECC Memory Supported
20	26	Max Resolution DP
10	11	Cache
10	12	Bus Speed
10	30	PCI Express Configurations
10	32	Temperature

Cluster Analysis

Cluster analysis was performed after the feature selection analysis as a complimentary evaluation for validating the selected attributes. The cluster objective function use iterations based on minimizing an objective function that represents the distance from any given data point to a cluster centre weighted by that data point's membership grade.

It is shown in this plot the times that each cluster reached the maximum of iterations, or when the objective function improvement between two consecutive iterations is less than the minimum amount of improvement specified. Results are depicted in Figure 5-38, were in part (a) shows the class interaction for the feature selected attributes thermal design power, recommended customer price and the response product collection; (b) presents the class partition between processor

base frequency, the target, and recommended customer price. These clusters found, confirm what the feature selection suggests, which is that the significance of thermal design power, and recommended customer price attributes against the response (product collection). On the other hand, the selected target for pattern recognition when interacting with a not significant attribute does not show significance.

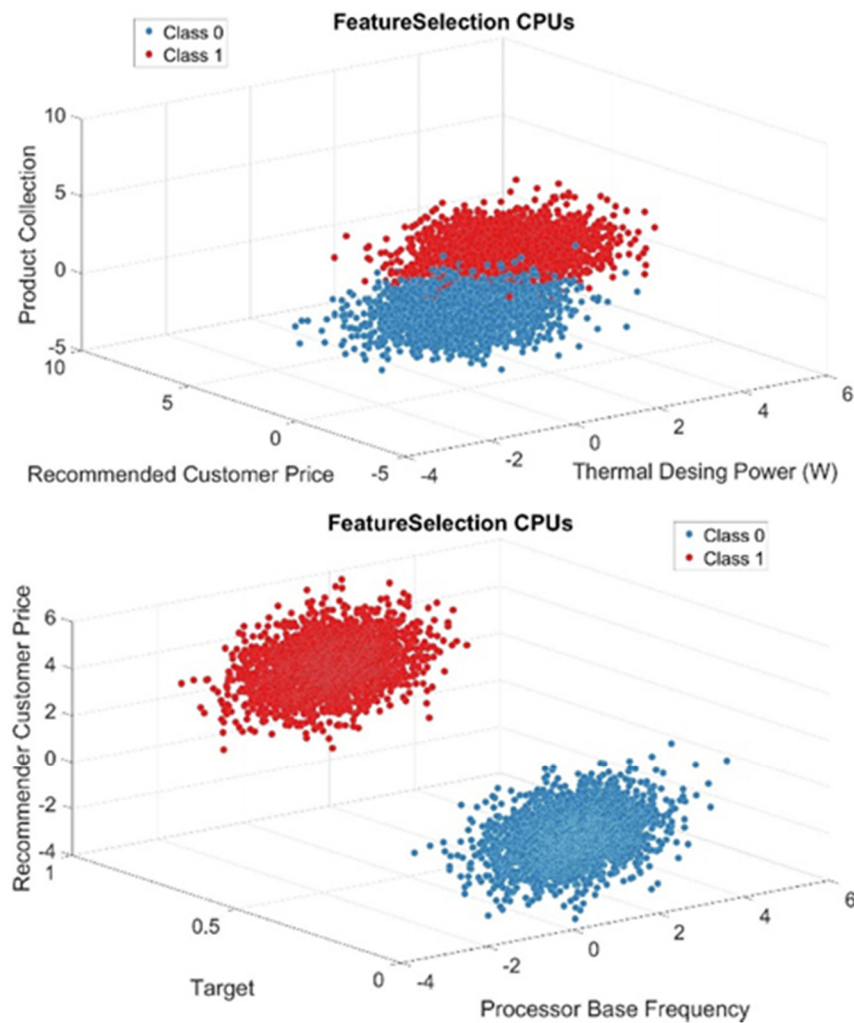


Figure 5-38 Clusters found for the CPUs dataset. Upper (a), lower (b).

Statistical Analysis

Finally, we proceed to validate the significance of the selected attributes using statistical test of the coefficient of determination (R^2). This test helps to determine if the used attributes were significant predictors. The interaction tested is depicted in Figure 5-39, showing customer recommended price, thermal design power, and max number of memory channels. The R^2 value obtained was 0.9574 after excluding some residual values as shown in figure (b), but without removing the residuals, the value scored was 0.8454. This test helped also to

detect the significance of other interactions that feature selection did not show, like considering the attribute max number of memory channels, leading to conclude that validation is always necessary, and data analysis can only be considered as a recommendation approach.

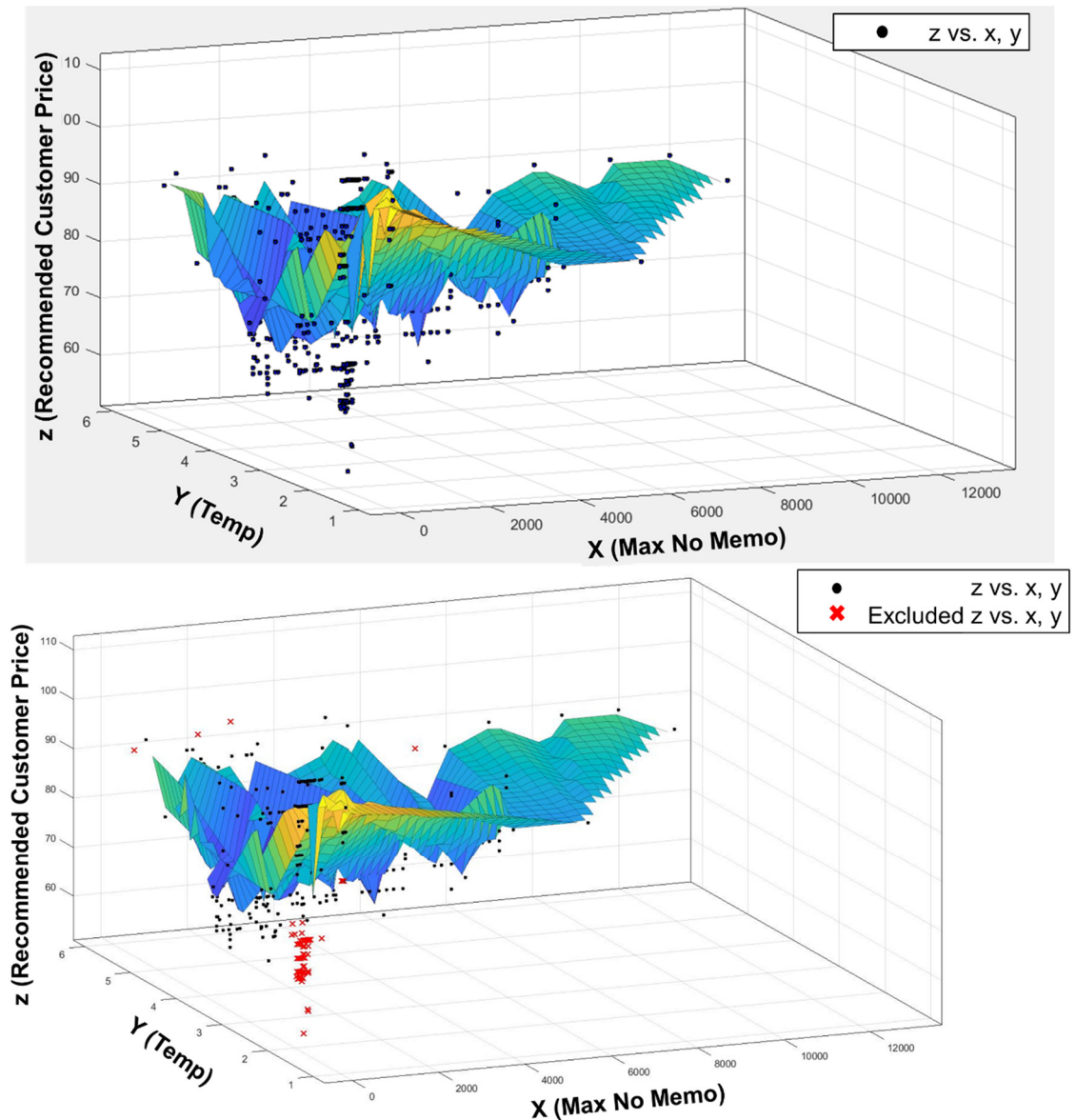


Figure 5-39 Surface plot for coefficient determination of predictive significant values.

The next section will present a comparison and evaluation of machine learning approaches for mining design attributes, where all the approaches applied in this chapter will be evaluated in terms of accuracy.

5.4 Evaluation of the Cases as a Result of Machine Learning Approaches

The evaluation of the used AI approaches for predicting customer needs and wants is presented in this section. In the previous sections were presented four case studies for which the results already lead to preliminary conclusions. Different to what is presented before, this chapter discusses which machine learning methods scored more accurate results. Therefore, the discussion presented here can help to make the final conclusions and annotations when predicting customer needs and wants for i4 and smart design. This section also represents the evaluation of the model for the training dataset, i.e. validate if the obtained model scores desired accuracy or predictive value against the original dataset.

With the inclusion of comparisons and evaluations, we aim at arriving at answers to the objectives stated in Chapter 1. Through these evaluations, we shall determine how both manufacturers and customers may benefit from such analysis and which methodologies lead to an accurate analysis.

5.4.1 Car Case Evaluation

The results obtained from this first stage when analysing the car evaluation dataset involved the accuracy of cluster and classification models. In Table 5–11 the accuracy comparison of the machine learning approaches when analysing the car evaluation dataset is presented.

Table 5–11 Model accuracy evaluation of AI approaches for the car evaluation dataset

Classifier	Accuracy %	Description
SOM	97.4%	Average clustering coefficient = 0.833. Training time: 21 sec. Categorical predictors: explain 97% of variance.
Simple means	k- 100%	1210 clusters were correctly classified into the unaccepted class. Training time: 28 sec. Categorical predictors: explain 100% of variance.
Ensemble bagged trees	90.9%	Prediction speed: 5700 obs/sec. Training time: 5.152 sec Categorical predictors: explain 90.9 % of variance.
SVM	77.1%	Kernel function: cubic Prediction speed: 11000 obs/sec. Training time: 8.3974 sec.

Constraint level box: 2
 Multiclass method: one vs one
 Categorical predictors: explain 77.1% of variance.

In the first evaluated dataset the simple k-means approach proved to be an effective method for pattern recognition, therefore unsupervised learning turned out to be more accurate. When comparing data mining techniques is necessary to have in mind what type of analysis is required, which in this case for the customized design it was necessary that the obtained model explained the variability of the phenomena involved. In this specific case the level of acceptance of car models when involved with other variables like buy price, repair price, door, person, size, and safety. Both SOM and simple k-means registered a longer training time than ensemble trees and SVM, but the accuracy does not reflect the same results.

5.4.2 Automobile Case Evaluation

For this case study it was implemented the fuzzy c-means clustering and also tested ensemble trees and SVM, but not anymore the SOM and simple k-means because of the combination of too many categorical instances. For this cases is suggested in [131] that when dealing with categorical values that do not represent numbers but enumerations (body style, manufacturer, engine type, etc.) is better to consider other methods that do not rely on Euclidian cost function that penalizes the performance or accuracy of the predictive model. Reason why we evaluated the aforementioned approaches presented in Table 5–12.

Table 5–12 Model accuracy evaluation of AI approaches for the automobile dataset

Classifier	Accuracy %	Description
Fuzzy c-means	84.4%	Prediction speed: 160 obs/sec Training time: 36.99 sec. Categorical predictors: explain 84.4% of variance.
Ensemble bagged trees	81.5%	Prediction speed: 550 obs/sec. Training time: 6.3474 sec Categorical predictors: explain 81.5% of variance.
SVM	80%	Kernel Function: cubic Prediction speed: 1200 obs/sec. Training time: 1.6843 sec. Constraint level box: 2 Multiclass method: one vs all Categorical predictors: explain 80.0% of variance.

The evaluation presented above shows how the combination of categorical and numerical instances for the automobile dataset required different techniques to obtain the predictive model and analysis. The fuzzy c-means approach reached the highest accuracy rate, but the training time was the longest. Ensemble trees, on the other hand, had a bit less accuracy percentage but significantly short training time. And finally, SVM's model performance was a bit short in terms of desirability with an accuracy of 80% and a training time of 1.6843 seconds, using the cubic kernel function.

5.4.3 Fuel Economy Case Evaluation

In Table 5–13, the accuracy of each classifier is listed. The dataset encompassed 52 attributes split into 23 categorical and 29 numerical ones. The total instances considered for this dataset was 4655.

Table 5–13 Model accuracy evaluation of AI approaches for the fuel economy dataset.

Classifier	Accuracy %	Description
Decision Tree	94.2%	Prediction speed: 30000 obs/sec Training time: 10.949 sec. Categorical predictors: explain 94.2% of variance.
SVM	14.3%	Kernel function: Cubic Prediction speed: 53000 obs/sec Training time: 39.563 sec Constraint level box: 2 Multiclass method: one vs all SVM was tested using several kernel functions apart from cubic, those include linear, quadratic, and fine Gaussian SVM. It was not able to explain most of the variance of the predictive model.
Ensemble bagged trees	99.2%	Prediction speed: 5500 obs/sec Training time: 19.159 sec. Categorical predictors: explain 99.2% of variance.

The accuracy evaluation of the machine learning techniques presented above lead to conclude that ensemble bagged trees performed excellently above the other tested approaches. Still, the training time reached with the ensemble bagged trees was not the shortest, but in terms of prediction is a good model. The decision trees also scored a good result, and in less time, but the problem that we are trying to solve involves prediction, therefore is better to maintain the most accurate model. Lastly, the results obtained from the SVM classification model

were poor, and even the time scored is the longest. For this specific case study is not recommended to use SVM classifier.

5.4.4 CPU Case Evaluation

In Table 5—14, the model accuracy of each classifier technique is listed. This dataset used to train the classification model encompassed 39 attributes split into 9 numerical values, and 30 categorical. The total observations considered in this dataset were 2298.

Table 5—14 Model accuracy evaluation of AI approaches for the CPU dataset.

Classifier	Accuracy %	Description
Ensemble Boosted Tree	58%	Prediction speed: 3600 obs/sec Training time: 34.416 sec. Categorical predictors: explain 58% of variance.
SVM	16.4%	Kernel function: Cubic Prediction speed: 14000 obs/sec Training time: 129.21 sec Constraint level box: 2 Multiclass method: one vs all SVM was tested using several kernel functions including linear, quadratic, cubic, and fine Gaussian SVM. It was not able to explain most of the variance of the predictive model.
Ensemble bagged trees	85%	Prediction speed: 2000 obs/sec Training time: 20.368 sec. Categorical predictors: explain 85% of variance.

This dataset or case study in specific involved a more complex process for classification, and the only machine learning technique capable of getting an accurate result, or at least one that was above the desired rate was the ensemble bagged trees. The ensemble bagged trees scored an accuracy value of 85% and a reasonable short training time. Then the ensemble boosted trees did not reach a desirable accurate value with 58% and this value cannot be used or is not recommended for prediction. The lowest value for accuracy was the SVM and also took the longest time, so again for this case study is not suitable to use SVM approaches.

5.5 Summary

From the results presented in this chapter, it can be concluded that many approaches tested are able to obtain satisfactory results of predicting customer

needs and wants. Of course, every single case study faces particular challenges to overcome, and different ways of analysing the inputs lead to improvements. The analysis presented in subsection 5.3.1 for the car evaluation shows good results in practice, but in this case study we did not focus on the visualization part. There exists room for improvement in regards of presenting as part of the analysis, where plots could actually help in the decision-making process. In the analysis presented in 5.3.2 for the automobile dataset, the implemented visual part as well could lead to a more intuitive analysis. Nonetheless, for the automobile dataset, a complete analysis has been performed with simple approaches, where feature selection analysis was not necessary since the desired results were already obtained.

The fuel economy dataset presented in section 5.3.3 has represented a bigger challenge, and part of the analysis there has involved evaluating several classification methods to test the effectiveness of each approach. The plots helped visualize the phenomena involved in this particular case, and because of this analysis, we were able to detect patterns and behaviours and obtain the desired prediction. We have found that fuzzy clustering complements well the analysis acquired, and both are useful if the case study or application involving many attributes to analyse.

For the last case study, in section 5.3.4, improvements have been considered. In this particular case, more complete analysis was obtained. The focus there was to achieve prediction, but as well to be able to recommend a concise number of attributes using feature selection in which both customers and designers would benefit. It was decided to include the statistical analysis, as part of the feature selection process, just to validate the accuracy of the results. This added robustness to the whole closed loop cycle, in terms of making the best decision when customizing a product according to individual needs.

In section 5.4 the evaluation results of machine learning techniques were presented. The evaluation consists of a comparison in model accuracy from the trained dataset against the original data, to determine if the obtained mathematic representation is suitable for use in prediction since one of the main objectives in this work was to predict customer needs and wants. Every single case study presents a specific challenge. It is seen that the performance of the classifier

mainly depends on the characteristics of the dataset. This is the reason why empirical tests need to be performed [92]. Therefore, in this chapter, we have presented the necessary tests to determine which classification models are more suitable for achieving prediction of customer needs and wants. As a result, it can be concluded that a common denominator for accurate results and performance along the case studies was found in the ensemble decision trees that always scored desired values. Although, simple k-means scored good values on prediction, this approach can only work with numerical data, as discussed in [133], where the mixture of attributes (categorical and numerical) needs a special treatment for the algorithm to code the sample data represented as discrete space and make a Euclidean distance representation to make it meaningful. Conversely, the SVM approach has never scored a desired percentage of accuracy. Different kernel functions were used for the SVM, and the cubic function presented the most accurate results for predictive models, but in practice the larger the attribute number was, the less accurate the model was. Thus, this technique is not recommended when dealing with a dataset that involves a mixture of categorical and numerical inputs, or where the dataset presents larger number of attributes. Approaches of SOM, cluster k-means, and fuzzy c-means have proved to be reliable when dealing with datasets that do not involve a high level of complexity. However, as discussed previously, when analysing data it is necessary to have a level of visualization, which none of these approaches provide properly.

Chapter 6 Conclusions and Future Work

This chapter presents conclusions and future work in 3 sections. The first section is about the discoveries obtained using the machine learning approaches and data analysis in general. The second section concludes the connection of the hypotheses stated in Chapter 3 to the obtained results, where the questions that correspond to the problem statement of Chapter 1 are also answered. Finally, in section 3, future directions are analysed.

6.1 General Conclusion

Machine learning for data-mining in this work has helped identify, predict, and recommend potential customer needs and wants, which manufacturers can consider as design elements for customizing products. The importance of this work lies in the need that current manufacturing has when moving to what is considered agile manufacturing. It is shown relatively efficient to obtain meaningful results from big data for mass customization. Using the perspective of i4 in this framework, we have developed a methodology that comprises multiple stages for addressing customer needs and wants and dealing with the gaps between the factories of today and the vision of i4-customized production.

This methodology has been tested in several applications as case studies, including consumer car evaluation, automotive vehicle characteristics, fuel economy, and computer parts. These case studies have helped us consolidate and validate the analysis. The following results have been obtained:

1. A classification approach has accurately predicted potential customer needs and wants, and this is achieved most consistently by the ensemble bagged trees.
2. Clustering analysis is able to identify partitioning and identification of patterns. The results reveal more specific significant attributes, which help narrow the features for design for agile manufacturing.
3. Intelligent search in the design process allows customers' needs and wants to be covered predictively. Virtual prototypes can hence be tuned

beforehand by customers when knowing the significant and predicted values obtained in the prediction model.

4. Considering the decision-making process, visualization helps the analysis be more appealing and intuitive. The plots presented in therein are not too complex to interpret and help accelerate decision making.
5. This way, manufacturers can make customer-oriented decisions using customer-driven informatics, design, AI-based recommended approaches and automation.
6. Data mining and data analytics help identify the influence of product characteristics, classification, attribute selection, clustering, and interpretation of customers' needs and wants.
7. It has been demonstrated that ensemble bagged trees and complex tree classifiers work well when trying to predict and select customers' needs and wants.
8. These analyses can contribute to manufacturing from the management perspective as an enabler of innovation according to customers' needs and wants and thus help companies avoid unnecessary product differentiation.

Conclusions concerning each dataset are detailed as follow.

6.1.1 Car Evaluation Dataset

1. SOM clustering reflects the attributes of the car as revealed in the case study, where the customer cares less about the "door" attribute.
2. The results also reveal that for car customization, "very good" and "good" cannot be easily met. Hence, it is predicted that the manufacturer should focus on the attributes on car sealing and on offerings of high-security and not on other attributes.

3. Simple k-means has been able to obtain a more accurate predictive model than other approaches do. For this specific case, this approach is seen reliable, although its visualization has presented a less complete analysis.

6.1.2 Automobile Dataset

1. In the case study, the results reveal that customer behaviour is based on 5 attributes (number-of-doors, drive-wheels, height, engine-type, number-of-cylinders).
2. Fuzzy c-means has performed a good partition on the dataset and has identified 3 clusters for classification.
3. Fuzzy c-means obtained the predictive model with a better percentage of accuracy. For practical implementation, this approach is relatively reliable and easy to use.

6.1.3 Fuel Economy Dataset

1. The model that accurately predicts customers' potential needs and wants has been obtained with ensemble-bagged trees. With this method, an accuracy of 99.3% was obtained.
2. For the fuel economy dataset, the results have confirmed that the method is working, i.e., if the customer wants to acquire a car in which fuel consumption is relatively low, then he/she should consider mini-compact cars based on the number of cylinders, gears, and type of drive (manual/automatic).
3. The car manufacturers that have presented misclassification to the predictive model of the fuel economy dataset are revealed as Audi, Bugatti, Chrysler Group, FCA Italy, Lamborghini, Mobility Ventures, Paganini Automobili, and Volkswagen.
4. For the clustering analysis, fuzzy c-means has performed a good partition and identification of three clusters, where multiple clustering approaches

were tested. Neither the simple k-means nor SOM could handle this challenge due to multiple variables or complexity of the datasets.

5. Through analysing this dataset, it is concluded that the ensemble bagged trees approach works better with complex datasets, and the fuzzy c-means works better for pattern identification for data analysis.

6.1.4 CPU Dataset

1. On the CPU dataset, the analysis has shown a recommended set of attributes that manufacturers can use to design a computer that reflects the customer's subconscious needs and wants. Significant features include system price, thermal power, memory types, thermal monitoring technologies, number of cores, and graphic base frequency, among many others.
2. Classification analysis has helped isolate the product collection Intel® Celeron® Processor J Series that has scored a misclassification, thus making it insignificant for the prediction model.
3. The classification approach that has accurately predicted customers' needs and wants is the ensemble bagged trees. The accuracy obtained with this method was 85%.

6.2 Reflections on the Hypotheses

Given the objectives stated in Chapter 1, this section answers the questions posed in the hypotheses of Chapter 3. Recapitulating about the questions for each hypothesis, conclusions are drawn as follow.

H1: It is possible to develop a framework capable of automatically predict the design attributes that best reflect what customers need and want in a product.

- Q1. How can a generalized framework be developed, which approaches can effectively predict the design attributes, and how to design smart products effectively to reflect what customers need and want in a product?

A1. In Chapter 4 the different stages of the proposed frameworks are presented. At each stage, the thesis has made different discoveries, challenges, and ways of addressing customer needs and wants. For this, different frameworks have been developed. The main focus was to develop a generalized framework able to automatically predict customer needs and wants. Consequently, turning customer needs and wants into design attributes for manufacturing a product. Through this work, we have discovered that it would be best to make predictions based on users' behaviour. Therefore, making easier selecting one setting instead of others, classifying the design attributes (based on the behaviour analysis already made), and finally recommending which set of attributes describe individual needs for a given product.

Results shown in chapter 5 lead to the conclusion that data mining techniques are suitable for predicting effectively design attributes. Moreover, in chapter 5, section 4 the evaluation results made from a combination of machine learning approaches proved to reflect desired conclusions when analysing the data. In specific ensemble trees, feature selection, and fuzzy clustering are effective approaches for classifying, recognizing patterns, and selecting features that best matched with customer needs and wants. Chapter 5 shows the results obtained when integrating computational intelligence. Efficiency needs to be measured accurately, and data mining techniques give the opportunity to know in specific how design reflects what customers need and want. Here Computer Automated Design plays a pivotal role, since smart products require constant development, and the framework proposed in chapter 4, section 1 can deal with automation and prediction by continuously evolving designs using AI and automating the process. Designs are improved from a digital platform that considers the analysis obtained from historical data.

H11: It is possible to obtain a model capable of accurately predict customer needs and wants for at least 85% of classified design attributes.

Q2. How can design attributes be used to make predictions, which AI approaches can be tested, and how can classification models be reliable when showing less than 85 %?

A1. Presented in chapter 3, section 3, design attributes are characteristic properties of a product, such is that in this work can be changed by an individual in order to fulfil his/her desires when customizing a product. This behaviour of changing, selecting, and customizing product designs can be classified, based on each individual configuration which in turn can provide insight of future events. Once this behaviour is modelled, is possible to match what customers would need and want in future events, because the design attributes are determined by each product and the way it is manufactured. Initially, it was decided to test machine learning approached used commonly for example, SOM, simple k-means, SVM, and decision trees for supervised learning as shown in chapters 5 and 6.

Moving forward with the complexity of different case studies, it was discovered that ensemble trees provide a more accurate representation of customer needs and wants. For every mathematical representation that tries to explain the given observations, considered as independent variables in a model, many indicators can be used to minimize the error when predicting possible values of the dependent variable. In this work, is included one case study analysis that shows these statistical indicators when validating prediction against the known observations. This error can be minimized once the used data is trained with sufficient information, allowing to make reliable predictions. It is desirable to use mathematical representations that present an accuracy above 95%, and literature suggests that validation is essential if is decided to use models that score any percentage below 95%.

HIII: It is possible to identify effective ways of achieving customization for i4 and smart manufacturing.

Q1. What are the identified challenges to be tackled, which methods are effective for achieving mass customization, and what particular ways does i4 deal with mass customization?

A1. Extracted from the literature review presented in chapter 2, section 5 shows that the challenges focus on business models, value-creation network, products, and processes. From here it was concluded that how data is managed inside a company can lead to effectively satisfy customer needs

and wants, since this is the main goal of customization. Therefore, we focused on tackling challenges that had to do with data analytics. i4 and smart manufacturing claim to address mass customization at mass production costs, but this challenge can never be achieved if a reliable analysis is made beforehand. In chapter 3, section 5 are given several examples of companies like YouTailor®, Bombsheller® and MyMuesli® where they offer through their website products that cannot be found in the store shelves, demonstrating that this is not a vision of the future, beyond that, is a necessity from the customers.

In many i4 demonstrations, manufacturers focus on the use of embedded systems interconnected to each other. The success of many current cases of mass customization relies on making available a virtual platform where the customer can interact with the design stage of their desired product. This interaction and selection are stored for future purchases, so the system can gain information about individual needs of users, and most important having models based on customers' behaviour. Chapter 2, first section presents how i4 and smart manufacturing deals with customization that is by making extensive use of the IoT, flexible process provided by CPS and cloud services that enable users to track the progress of their order. Many companies in the last decade proved that customization is possible, but doing it massively requires to overcome the aforementioned challenges.

In this work, it has been highlighted the importance of customization in the coming 4th Industrial Revolution. A solid framework has been proposed that integrates most of the principles of smart technologies to realize i4. Industry 4.0 is characterized by bringing the innovation to the shop floor, and the key aspect for this is digitalization, where product design plays a decisive role. It has been discovered that in this stage designs can be customized according to individual needs without sacrificing manufacturing time and effort.

6.3 Future Directions

While the thesis has focused on predicting potential customer needs and wants for agile design and manufacture in an Industry 4.0 environment, future work will include integrating affective design approaches to a fully integration of

customers' sentiment about product attributes. The affective design approach can bring a more clear analysis and identification of customer needs and wants because of the integration of sentiment of design elements to the whole value chain and therefore, have a direct indicator of how efficient the model can be compared to the levels of affection a customer have towards design attributes or elements. Intelligence on customers' feelings can be coded into design elements to reduce misunderstanding and make predictions more accurate, which is complementary in point 6 of the general conclusions of this chapter. This approach could require the development of a questionnaire or survey, as targeted questions about individual feelings can improve the mining of customer needs and wants.

Further, the prediction may be validated and integrated by using virtual or augmented reality to collect more data in real time or to perform an exploratory test and train an algorithm with individual sentiment about perceived product characteristics, helping to improve point number 4 in the general conclusions in this chapter, i.e. decision making in real-time. In this regard, descriptive statistics may be integrated to facilitate the analysis and further improvements. Including more digitally aided technologies can also lead to improvements in, and adjustments of, product designs. Thus, this facilitates the process of an enhanced customization of products in real time.

Ways of measuring customer satisfaction are also a future direction, to help extending point number 3 in the general conclusions presented in this chapter. Retrieving such measurements can be used as indicators for manufacturers and businesses to customize their products more individually, a prediction model can be obtained easily when considering an indicator of customer needs and wants, in terms of weight attributes for the developed model.

It was discovered in chapter 5 that simple k-means can be useful when performing cluster analysis to numerical values, but not when dealing with a mixture of categorical and numerical values. A way forward can be exploring other k-means algorithms suitable for mixed attributes to see if are more effective than decision trees approaches.

SVM approaches can also be explored with different kernel functions, since the common cubic, linear, quadratic, and fine Gaussian functions were not effective

when dealing with datasets that involve a mixture of categorical and numerical instances and as well for the response. A way forward can to this can be trying with different kernel functions like Radial Basis Function or algorithms capable of dealing with canonical correlation analysis to replace features or predictors to obtain better prediction or know where to adjust the model.

References

1. Coetteleer, M., J. Holdowsky, and M. Mahto, *Additive Manufacturing paths to performance, innovation, and growth*. Deloitte Review, 2014. 1(19): p. 32.
2. Prinz, C., et al., *Learning Factory Modules for Smart Factories in Industrie 4.0*. Procedia CIRP, 2016. 54: p. 113-118.
3. Kagermann, H., W. Wahlster, and J. Helbig, *Recommendations for implementing the strategic initiative INDUSTRIE 4.0*. ACATECH NATIONAL ACADEMY OF SCIENCE AND ENGINEERING, 2013(6).
4. Xu, Y., G. Chen, and J. Zheng, *An integrated solution—KAGFM for mass customization in customer-oriented product design under cloud manufacturing environment*. The International Journal of Advanced Manufacturing Technology, 2016. 84(1): p. 85-101.
5. Franke, N., M. Schreier, and U. Kaiser, *The “I Designed It Myself” Effect in Mass Customization*. Management Science, 2010. 56(1): p. 125-140.
6. Fogliatto, F.S., G.J.C. da Silveira, and D. Borenstein, *The mass customization decade: An updated review of the literature*. International Journal of Production Economics, 2012. 138(1): p. 14-25.
7. Flores Saldivar, A.A., et al. *Industry 4.0 with Cyber-Physical Integration: A Design and Manufacture Perspective*. in *International Conference on Automation & Computing*. 2015. University of Strathclyde, Glasgow: IEEE.
8. Schneider, P., *Managerial challenges of Industry 4.0: an empirically backed research agenda for a nascent field*. Review of Managerial Science, 2018.
9. Schönemann, M., et al., *Simulation of matrix-structured manufacturing systems*. Journal of Manufacturing Systems, 2015(7).
10. Lasi, H., et al., *Industry 4.0*. Business & Information Systems Engineering, 2014. 6(42): p. 239-242.
11. MacDougall, W., *INDUSTRIE 4.0 SMART MANUFACTURING FOR THE FUTURE*. MECHANICAL & ELECTRONIC TECHNOLOGIES, GERMANY TRADE & INVEST, 2014(3): p. 40.
12. VDE-DKE GERMAN ASSOCIATION FOR ELECTRICAL, E.I.T., *the German Standardization Roadmap Industrie 4.0*. DKE STANDARDIZATION ROADMAP, 2014: p. 60.
13. Schlick, J., et al., *Industrie 4.0 in der praktischen Anwendung*, in *Industrie 4.0 in Produktion, Automatisierung und Logistik*, T. Bauernhansl, M. ten Hompel, and B. Vogel-Heuser, Editors. 2014, Springer Fachmedien Wiesbaden. p. 57-84.
14. Kagermann, H., *Change Through Digitization—Value Creation in the Age of Industry 4.0*, in *Management of Permanent Change*, H. Albach, et al., Editors. 2015, Springer Fachmedien Wiesbaden. p. 23-45.
15. Laili, Y., et al., *A Ranking Chaos Algorithm for dual scheduling of cloud service and computing resource in private cloud*. Computers in Industry, 2013. 64(5): p. 448-463.
16. Lee, J., H.-A. Kao, and S. Yang. *Service Innovation and Smart Analytics for Industry 4.0 and Big Data Environment*. in *Procedia CIRP*. 2014.
17. Stock, T. and G. Seliger, *Opportunities of Sustainable Manufacturing in Industry 4.0*. Procedia CIRP, 2016. 40: p. 536-541.
18. Akusok, A., et al., *Per-sample prediction intervals for extreme learning machines*. International Journal of Machine Learning and Cybernetics, 2018.

19. Kang, H.S., et al., *Smart manufacturing: Past research, present findings, and future directions*. International Journal of Precision Engineering and Manufacturing-Green Technology, 2016. **3**(1): p. 111-128.
20. Lucke, D., C. Constantinescu, and E. Westkämper, *Smart Factory - A Step towards the Next Generation of Manufacturing*, in *Manufacturing Systems and Technologies for the New Frontier*, M. Mitsuishi, K. Ueda, and F. Kimura, Editors. 2008, Springer London. p. 115-118.
21. Weiser, M., *The computer for the 21st century*. SIGMOBILE Mob. Comput. Commun. Rev., 1999. **3**(3): p. 3-11.
22. Wang, S., et al., *Implementing Smart Factory of Industrie 4.0: An Outlook*. International Journal of Distributed Sensor Networks, 2015(37).
23. Möller, D.P.F., *Digital Manufacturing/Industry 4.0*, in *Guide to Computing Fundamentals in Cyber-Physical Systems: Concepts, Design Methods, and Applications*. 2016, Springer International Publishing: Cham. p. 307-375.
24. Tseng, M.M. and J. Jiao, *Mass Customization*, in *Handbook of Industrial Engineering*. 2007, John Wiley & Sons, Inc. p. 684-709.
25. Yang, B. and N. Burns, *Implications of postponement for the supply chain*. International Journal of Production Research, 2003. **41**(9): p. 2075-2090.
26. Yun Li, K.H.A., Gregory C.Y. Chong, Wenyuan Feng, Kay Chen Tan, Hiroshi Kashiwagi, *CAutoCSD-Evolutionary Search and Optimisation Enabled Computer Automated Control System Design*. International Journal of Automation and Computing, 2004. **1**(17): p. 76-88.
27. Flores Saldivar, A.A., et al. *Self-organizing tool for smart design with predictive customer needs and wants to realize Industry 4.0*. in *World Congress on Computational Intelligence*. 2016. Vancouver, Canada: IEEE.
28. Wan, J., et al., *Mobile Services for Customization Manufacturing Systems: An Example of Industry 4.0*. IEEE Access, 2016. **4**: p. 8977-8986.
29. Pollard, D., S. Chuo, and B. Lee, *Strategies for mass customization*. Journal of Business & Economics Research, 2008. **6**(7): p. 77-86.
30. Schuh, G., et al., *Collaboration Mechanisms to Increase Productivity in the Context of Industrie 4.0*. Procedia CIRP, 2014. **19**(0): p. 51-56.
31. Schlaepfer, R.C. and M. Koch, *Challenges and solutions for the digital transformation and use of exponential technologies*. Industry 4.0 Deloitte, 2014. **1**(10): p. 32.
32. Porter, M.E., *The Competitive Advantage: Creating and Sustaining Superior Performance*, ed. N.F. Press. Vol. 1. 1985, NY, U. S. A: The Free Press.
33. Feller Andrew, D.D. Shunk, and T.D. Callarman, *Value Chains Versus Supply Chains*. BPT Trends, 2006. **March 2006**(12): p. 1-7.
34. Groover, M.P., *Automation, Production Systems, and Computer-integrated Manufacturing*. 2001 ed. 2001: Prentice Hall. 856.
35. Groover, M.P., *Single-station manufacturing cells, Automation, production systems, and computer-integrated manufacturing*, ed. C. Technologies. Vol. 2. 2016: Cram101. 88.
36. Groover, M.P., *Group technology and cellular manufacturing, Automation, production systems, and computer-integrated manufacturing*. 2007: Prentice Hall Press. 840.
37. ALAVUDEEN, A. and N. VENKATESHWARAN, *COMPUTER INTEGRATED MANUFACTURING*. Vol. 2. 2008, New Delhi: PHI Learning. 440.
38. Koren, Y., et al., *Reconfigurable Manufacturing Systems*. CIRP Annals - Manufacturing Technology, 1999. **48**(2): p. 527-540.

39. Qin, J., Y. Liu, and R. Grosvenor, *A Categorical Framework of Manufacturing for Industry 4.0 and Beyond*. Procedia CIRP, 2016. 52: p. 173-178.
40. Monostori, L., *Cyber-physical Production Systems: Roots, Expectations and R&D Challenges*. Procedia CIRP, 2014. 17(20): p. 9-13.
41. Technology, N.I.o.S.a., *Strategic R&D Oportunities for 21st Century Cyber-Physical Systems*. Foundations for Innovation in Cyber-Physical Systems Workshop, 2013(21): p. 32.
42. Lichen, L., *Model Integration and Model Transformation Approach for Multi-Paradigm Cyber Physical System Development*, in *Progress in Systems Engineering*, H. Selvaraj, D. Zydek, and G. Chmaj, Editors. 2015, Springer International Publishing. p. 629-635.
43. Lee, E.A. *Cyber Physical Systems: Design Challenges*. in *Object Oriented Real-Time Distributed Computing (ISORC), 2008 11th IEEE International Symposium on*. 2008.
44. Günthner, W., E. Klenk, and P. Tenerowicz-Wirth, *Adaptive Logistiksysteme als Wegbereiter der Industrie 4.0*, in *Industrie 4.0 in Produktion, Automatisierung und Logistik*, T. Bauernhansl, M. ten Hompel, and B. Vogel-Heuser, Editors. 2014, Springer Fachmedien Wiesbaden. p. 297-323.
45. Sztipanovits, J., et al., *OpenMETA: A Model- and Component-Based Design Tool Chain for Cyber-Physical Systems*, in *From Programs to Systems. The Systems perspective in Computing*, S. Bensalem, Y. Lakhneck, and A. Legay, Editors. 2014, Springer Berlin Heidelberg. p. 235-248.
46. Wolf, M., *Chapter 1 - Embedded Computing*, in *High-Performance Embedded Computing (Second Edition)*, M. Wolf, Editor. 2014, Morgan Kaufmann: Boston. p. 1-58.
47. Kaffka, G., *Transitioning embedded systems to intelligent environments – A journey through evolving technologies*, Satwant Kaur. Technological Forecasting and Social Change, 2015. 90, Part B(0): p. 651-652.
48. Lee, E.A. and S.A. Seshia, *Introduction to Embedded Systems: A Cyber-Physical Systems Approach*. 2016: The MIT Press. 568.
49. Lee, J., et al., *Recent advances and trends in predictive manufacturing systems in big data environment*. Manufacturing Letters, 2013. 1(45): p. 38-41.
50. Lee, J., B. Bagheri, and H.-A. Kao, *A Cyber-Physical Systems architecture for Industry 4.0-based manufacturing systems*. Manufacturing Letters, 2015. 3: p. 18-23.
51. Nauck, D., et al., *Predictive Customer Analytics and Real-Time Business Intelligence*, in *Service Chain Management*, C. Voudouris, D. Lesaint, and G. Owusu, Editors. 2008, Springer Berlin Heidelberg. p. 205-214.
52. Simmhan, Y. and S. Perera, *Big Data Analytics Platforms for Real-Time Applications in IoT*, in *Big Data Analytics: Methods and Applications*, S. Pyne, B.L.S.P. Rao, and S.B. Rao, Editors. 2016, Springer India: New Delhi. p. 115-135.
53. Rabiner, L. and B.H. Juang, *An introduction to hidden Markov models*. ASSP Magazine, IEEE, 1986. 3(12): p. 4-16.
54. Heckerman, D. and M.P. Wellman, *Bayesian networks*. Commun. ACM, 1995. 38(13): p. 27-30.
55. Adolph, M., *Big Data: Big today, normal tomorrow*. ITU-T Technology Watch Report, 2013(28): p. 28.

56. Gubbi, J., et al., *Internet of Things (IoT): A vision, architectural elements, and future directions*. Future Generation Computer Systems, 2013. **29**(7): p. 1645-1660.
57. Bohlouli, M., et al., *Towards an Integrated Platform for Big Data Analysis*, in *Integration of Practice-Oriented Knowledge Technology: Trends and Prospectives*, M. Fathi, Editor. 2013, Springer Berlin Heidelberg. p. 47-56.
58. Singh, D. and C.K. Reddy, *A survey on platforms for big data analytics*. Journal of Big Data, 2014. **2**(1): p. 8.
59. Tien, J.M., *The next industrial revolution: Integrated services and goods*. Journal of Systems Science and Systems Engineering, 2012. **21**(3): p. 257-296.
60. Kull, H., *Introduction*, in *Mass Customization: Opportunities, Methods, and Challenges for Manufacturers*. 2015, Apress: Berkeley, CA. p. 1-6.
61. Kull, H., *Intelligent Manufacturing Technologies*, in *Mass Customization: Opportunities, Methods, and Challenges for Manufacturers*. 2015, Apress: Berkeley, CA. p. 9-20.
62. Flores Saldivar, A.A., et al. *Identifying Smart Design Attributes for Industry 4.0 Customization Using a Clustering Genetic Algorithm*. in *International Conference on Automation & Computing*. 2016. University of Essex, Colchester city, UK: IEEE.
63. Isaacson, W. *The real leadership lessons of Steve Jobs*. Harvard Business Review, 2012. **4**, 92-102.
64. Yang, S.L. and T.F. Li, *Agility evaluation of mass customization product manufacturing*. Journal of Materials Processing Technology, 2002. **129**(1): p. 640-644.
65. Um, J., *The impact of supply chain agility on business performance in a high level customization environment*. Operations Management Research, 2017. **10**(1): p. 10-19.
66. Wolpert, D.H., *The Lack of A Priori Distinctions Between Learning Algorithms*. Neural Computation, 1996. **8**(7): p. 1341-1390.
67. Guszczka, J., H. Lewis, and P. Evans-Greenwood, *Cognitive collaboration Why humans and computers think better together*. Deloitte Review, 2017. **1**(20): p. 7-30.
68. Tien, J.M., *Internet of connected ServGoods: Considerations, consequences and concerns*. Journal of Systems Science and Systems Engineering, 2015. **24**(2): p. 130-167.
69. Tien, J.M., *Internet of Things, Real-Time Decision Making, and Artificial Intelligence*. Annals of Data Science, 2017. **4**(2): p. 149-178.
70. Tien, J.M., *Toward a decision informatics paradigm: a real-time, information-based approach to decision making*. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2003. **33**(1): p. 102-113.
71. Ji-Hyeong, H. and C. Su-Young. *Consideration of manufacturing data to apply machine learning methods for predictive manufacturing*. in *2016 Eighth International Conference on Ubiquitous and Future Networks (ICUFN)*. 2016.
72. Tiihonen, J. and A. Felfernig, *An introduction to personalization and mass customization*. Journal of Intelligent Information Systems, 2017. **49**(1): p. 1-7.
73. Jannach, D., et al., *Recommender Systems: An Introduction*. 2010, Cambridge: Cambridge University Press.

74. Mobasher, B., *Data mining for web personalization*, in *The adaptive web*, B. Peter, K. Alfred, and N. Wolfgang, Editors. 2007, Springer-Verlag. p. 90-135.
75. Linden, G., B. Smith, and J. York, *Amazon.com recommendations: item-to-item collaborative filtering*. IEEE Internet Computing, 2003. 7(1): p. 76-80.
76. Chee, S.H.S., J. Han, and K. Wang. *RecTree: An Efficient Collaborative Filtering Method*. 2001. Berlin, Heidelberg: Springer Berlin Heidelberg.
77. Konstan, J.A., et al., *GroupLens: applying collaborative filtering to Usenet news*. Commun. ACM, 1997. 40(3): p. 77-87.
78. Stormer, H., *Improving product configurators by means of a collaborative recommender system*. International Journal of Mass Customisation, 2009. 3(2): p. 165-178.
79. Adomavicius, G. and A. Tuzhilin, *Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions*. IEEE Transactions on Knowledge and Data Engineering, 2005. 17(6): p. 734-749.
80. Schafer, J.B., et al., *Collaborative Filtering Recommender Systems*, in *The Adaptive Web: Methods and Strategies of Web Personalization*, P. Brusilovsky, A. Kobsa, and W. Nejdl, Editors. 2007, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 291-324.
81. Mavridou, E., et al., *Mining affective needs of automotive industry customers for building a mass-customization recommender system*. Journal of Intelligent Manufacturing, 2013. 24(2): p. 251-265.
82. Xu, Z., V. Sugumaran, and N.Y. Yen, *Special issue: algorithmic and knowledge-based approaches to assessing consumer sentiment in electronic commerce*. Electronic Commerce Research, 2018. 18(1): p. 1-1.
83. Thompson, C.A., et al., *A personalized system for conversational recommendations*. J. Artif. Int. Res., 2004. 21(1): p. 393-428.
84. Zhang, Y. and J. Jiao, *An associative classification-based recommendation system for personalization in B2C e-commerce applications*. Expert Systems with Applications, 2007. 33(2): p. 357-367.
85. Ulz, T., et al., *Human computation for constraint-based recommenders*. Journal of Intelligent Information Systems, 2017. 49(1): p. 37-57.
86. Grosso, C., C. Forza, and A. Trentin, *Supporting the social dimension of shopping for personalized products through online sales configurators*. Journal of Intelligent Information Systems, 2017. 49(1): p. 9-35.
87. Chang, W.-c. and T.Y. Wu, *Exploring Types and Characteristics of Product Forms*. 2007. 2007.
88. Risdiyono and P. Koomsap, *Design by customer: concept and applications*. Journal of Intelligent Manufacturing, 2013. 24(2): p. 295-311.
89. Khalid, H.M., et al., *Elicitation and analysis of affective needs in vehicle design*. Theoretical Issues in Ergonomics Science, 2012. 13(3): p. 318-334.
90. Nagamachi, M., *Kansei engineering as a powerful consumer-oriented technology for product development*. Applied Ergonomics, 2002. 33(3): p. 289-294.
91. Jiao, R.J., et al., *Analytical affective design with ambient intelligence for mass customization and personalization*. International Journal of Flexible Manufacturing Systems, 2007. 19(4): p. 570-595.
92. Witten, I.H. and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. 2000: Morgan Kaufmann.
93. Kingston, G.B., H.R. Maier, and M.F. Lambert, *A probabilistic method for assisting knowledge extraction from artificial neural networks used for*

- hydrological prediction*. Mathematical and Computer Modelling, 2006. 44(5-6): p. 499-512.
94. Bezdek, J.C., R. Ehrlich, and W. Full, *FCM: The fuzzy c-means clustering algorithm*. Computers & Geosciences, 1984. 10(2): p. 191-203.
 95. Bezdek, J.C., *Objective Function Clustering*, in *Pattern Recognition with Fuzzy Objective Function Algorithms*. 1981, Springer US: Boston, MA. p. 43-93.
 96. Bougoudis, I., L. Iliadis, and S. Spartalis, *Comparison of Self Organizing Maps Clustering with Supervised Classification for Air Pollution Data Sets*, in *Artificial Intelligence Applications and Innovations: 10th IFIP WG 12.5 International Conference, AIAI 2014, Rhodes, Greece, September 19-21, 2014. Proceedings*, L. Iliadis, I. Maglogiannis, and H. Papadopoulos, Editors. 2014, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 424-435.
 97. Ludwig, S.A., *MapReduce-based fuzzy c-means clustering algorithm: implementation and scalability*. International Journal of Machine Learning and Cybernetics, 2015. 6(6): p. 923-934.
 98. Bezdek, J.C., *Pattern Recognition with Fuzzy Objective Function Algorithms*. 1981: Kluwer Academic Publishers. 256.
 99. Dietterich, T.G., *Approximate statistical tests for comparing supervised classification learning algorithms*. Neural Comput., 1998. 10(7): p. 1895-1923.
 100. Breiman, L., et al., *Classification and Regression Trees*. 1984: Taylor & Francis.
 101. Brydon, M. and A. Gemino, *Classification trees and decision-analytic feedforward control: a case study from the video game industry*. Data Mining and Knowledge Discovery, 2008. 17(2): p. 317-342.
 102. Canzian, L., Y. Zhang, and M.v.d. Schaar, *Ensemble of distributed learners for online classification of dynamic data streams*. IEEE Transactions on Signal and Information Processing over Networks, 2015. 1(3): p. 180-194.
 103. Tekin, C., J. Yoon, and M.v.d. Schaar, *Adaptive Ensemble Learning With Confidence Bounds*. IEEE Transactions on Signal Processing, 2017. 65(4): p. 888-903.
 104. Yoon, J., W.R. Zame, and M.v.d. Schaar, *ToPs: Ensemble Learning With Trees of Predictors*. IEEE Transactions on Signal Processing, 2018. 66(8): p. 2141-2152.
 105. *Support Vector Machines in Classification and Regression – An Introduction*, in *Kernel Based Algorithms for Mining Huge Data Sets: Supervised, Semi-supervised, and Unsupervised Learning*. 2006, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 11-60.
 106. Dai, W., H. Li, and Q. Liu, *Application Research of Support Vector Machine Classification Algorithm*, in *Unifying Electrical Engineering and Electronics Engineering: Proceedings of the 2012 International Conference on Electrical and Electronics Engineering*, S. Xing, et al., Editors. 2014, Springer New York: New York, NY. p. 2103-2110.
 107. Liu, H. and H. Motoda, *Less Is More*, in *Feature Extraction, Construction and Selection: A Data Mining Perspective*, H. Liu and H. Motoda, Editors. 1998, Springer US: Boston, MA. p. 3-12.
 108. Salcedo-Sanz, S., et al., *Enhancing genetic feature selection through restricted search and Walsh analysis*. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2004. 34(4): p. 398-406.

109. Alexandre, E., et al., *Feature Selection for Sound Classification in Hearing Aids Through Restricted Search Driven by Genetic Algorithms*. IEEE Transactions on Audio, Speech, and Language Processing, 2007. 15(8): p. 2249-2256.
110. Kohavi, R. and G.H. John, *Wrappers for feature subset selection*. Artificial Intelligence, 1997. 97(1): p. 273-324.
111. Blum, A.L. and P. Langley, *Selection of relevant features and examples in machine learning*. Artificial Intelligence, 1997. 97(1): p. 245-271.
112. Weston, J., et al., *Feature selection for SVMs*, in *Proceedings of the 13th International Conference on Neural Information Processing Systems*. 2000, MIT Press: Denver, CO. p. 647-653.
113. Bradley, P.S., U.M. Fayyad, and O.L. Mangasarian, *Mathematical Programming for Data Mining: Formulations and Challenges*. INFORMS J. on Computing, 1999. 11(3): p. 217-238.
114. Yang, J. and V. Honavar, *Feature Subset Selection Using a Genetic Algorithm*, in *Feature Extraction, Construction and Selection: A Data Mining Perspective*, H. Liu and H. Motoda, Editors. 1998, Springer US: Boston, MA. p. 117-136.
115. Camarinha-Matos, L.M. and H. Afsarmanesh, *Collaborative systems for smart environments: Trends and challenges*, in *15th IFIP WG 5.5 Working Conference on Virtual Enterprises, PRO-VE 2014*, H. Afsarmanesh and L.M. Camarinha-Matos, Editors. 2014, Springer New York LLC. p. 3-15.
116. Azvine, B., et al. *Real Time Business Intelligence for the Adaptive Enterprise*. in *E-Commerce Technology, 2006. The 8th IEEE International Conference on and Enterprise Computing, E-Commerce, and E-Services, The 3rd IEEE International Conference on*. 2006.
117. Guszczka, J., *Smarter together: Why artificial intelligence needs human-centered design*. Deloitte Review, 2018. 1(22): p. 36-45.
118. Buhr, D. *Social Innovation Policy for Industry 4.0*. 2015. 3-24.
119. You, K.-Y. and L.-H. Xie, *Survey of Recent Progress in Networked Control Systems*. Acta Automatica Sinica, 2013. 39(38): p. 101-117.
120. Beckhoff, I.C. *Collaboration Accelerates the Internet of Things and Industry 4.0*. Intel Corporation technical report, 2015. 1-5 DOI: 332096-001US.
121. Brambley, S., *Industry 4.0: What is the benefit to the customer?*, in *Drives & Controls*, GAMBICA, Editor. 2015, GAMBICA: United Kingdom. p. 1.
122. Ang, J., et al., *Energy-Efficient Through-Life Smart Design, Manufacturing and Operation of Ships in an Industry 4.0 Environment*. Energies, 2017. 10(5): p. 610.
123. Saldivar, A.A.F., et al. *Attribute identification and predictive customisation using fuzzy clustering and genetic search for Industry 4.0 environments*. in *2016 10th International Conference on Software, Knowledge, Information Management & Applications (SKIMA)*. 2016.
124. Lichman, M., *UCI Machine Learning Repository*, U.o. California, Editor. 2013, School of Information and Computer Science.: Irvine, CA: University of California, School of Information and Computer Science.
125. Schlimmer, J.C., *Automonile Data Set*, W.s.A. Yearbook, Editor. 1985, UCI Machine Learning Repository: United States of America.
126. Energy, U.S.D.o., *Fuel Economy*, in *Source for fuel economy information*, U.S. Government, Editor. 1992, Office of Transportation & Air Quality: U. S. A. p. 8.

127. Kuhn, M. and K. Johnson, *A Short Tour of the Predictive Modeling Process*, in *Applied Predictive Modeling*. 2013, Springer New York: New York, NY. p. 19-26.
128. Batten, C., et al., *CPU DB*, in *Intel computer parts*, S. University, Editor. 2012, Stanford University's VLSI Research Group: U. S. A.
129. Shuai, S., Z. Laibin, and L. Wei. *Condition monitoring and fault diagnosis of rolling element bearings based on wavelet energy entropy and SOM*. in *Quality, Reliability, Risk, Maintenance, and Safety Engineering (ICQR2MSE)*, 2012 International Conference on. 2012.
130. Zhu, L. and Y. Yang. *Improvement of Decision Tree ID3 Algorithm*. 2017. Cham: Springer International Publishing.
131. Sharma Ritu (Sachdeva, Alam Afshar M., and R. Anita, *K-Means Clustering in Spatial Data Mining using Weka Interface*. *IJCA Proceedings on International Conference on Advances in Communication and Computing Technologies 2012*, 2012. *ICACACT*(1): p. 26-30.
132. Arcuri, A. and G. Fraser, *Parameter tuning or default values? An empirical investigation in search-based software engineering*. *Empirical Software Engineering*, 2013. **18**(3): p. 594-623.
133. Huang, Z., *Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values*. *Data Mining and Knowledge Discovery*, 1998. **2**(3): p. 283-304.